

MIXED SIGNAL  
VLSI  
**WIRELESS**  
*Circuits and*

---

---

**MIXED SIGNAL  
VLSI  
WIRELESS DESIGN**

**Circuits and Systems**

---

---

**MIXED SIGNAL  
VLSI  
WIRELESS DESIGN**

**Circuits and Systems**

**Emad N. Farag**  
*Lucent Technologies*

and

**Mohamed I. Elmasry**  
*University of Waterloo*

KLUWER ACADEMIC PUBLISHERS  
NEW YORK, BOSTON, DORDRECHT, LONDON, MOSCOW

eBook ISBN: 0-306-47307-0  
Print ISBN: 0-792-38687-6

©2002 Kluwer Academic Publishers  
New York, Boston, Dordrecht, London, Moscow

All rights reserved

No part of this eBook may be reproduced or transmitted in any form or by any means, electronic, mechanical, recording, or otherwise, without written consent from the Publisher

Created in the United States of America

Visit Kluwer Online at: <http://www.kluweronline.com>  
and Kluwer's eBookstore at: <http://www.ebooks.kluweronline.com>

# Contents

List of Figures	ix
List of Tables	xvii
Preface	xix
Acknowledgments	xxi
List of Acronyms	xxiii
<b>1. INTRODUCTION</b>	1
1.1 The History of Wireless Communications	1
1.2 Wireless Communications Basics	4
1.3 Wireless Communications Standards	9
1.4 Terminology	18
1.5 Book Organization	20
<b>2. WIRELESS COMMUNICATION SYSTEMS - OVERVIEW</b>	23
2.1 Introduction	23
2.2 Digital Communication Systems	23
2.3 Minimum Bandwidth Requirement (The Nyquist Limit)	27
2.4 The Shannon Limit	36
2.5 Optimum Receiver Design for Band-pass Digital Systems	45
<b>3. THE MOBILE RADIO</b>	55
3.1 Introduction	55
3.2 The Cellular Concept	55
3.3 Channel Impairments	66
3.4 Large Scale Fading	71
3.5 Doppler Shift	72
3.6 Small Scale Fading	75
3.7 Diversity Techniques in Wireless Communications	85
<b>4. DIGITAL MODULATION SCHEMES</b>	87
4.1 Introduction	87

4.2	Amplitude Shift Keying	88
4.3	Phase Shift Keying	91
4.4	Frequency Shift Keying	106
4.5	Digital Modulation Techniques Comparisons	111
5.	SPREAD SPECTRUM	115
5.1	Introduction	115
5.2	Basic Principles of Spread Spectrum	117
5.3	Spread Spectrum Techniques	118
5.4	Pseudorandom Noise Sequence	121
5.5	Practical Considerations in Spread Spectrum Systems	126
6.	RECEIVER ARCHITECTURES	133
6.1	Introduction	133
6.2	Noise Figure	133
6.3	Intermodulation distortion	138
6.4	The Superheterodyne Receiver	144
6.5	The Homodyne Receiver	150
6.6	Software Radio	151
7.	ANALOG TO DIGITAL CONVERSION	155
7.1	Introduction	155
7.2	Performance Metrics of Analog-to-Digital Converters	158
7.3	Sampling	158
7.4	Band-pass Sampling	165
7.5	Quantization	167
7.6	Types of Analog-to-Digital Converters	171
7.7	Sigma-Delta Analog-to-Digital Converters	188
8.	VLSI DESIGN ISSUES IN WIRELESS TRANSCEIVER DESIGN	195
8.1	Introduction	195
8.2	Transceiver Design Constraints	197
8.3	Baseband Subsystem Design	199
8.4	RF Subsystem Design	204
9.	LOW-POWER DESIGN TECHNIQUES	211
9.1	Introduction	211
9.2	Sources of Power Dissipation	213
9.3	Estimating the Power Dissipation	215
9.4	Low-Power Examples of Portable Systems	217
9.5	Reducing the Power Dissipation at the Device and Circuit Levels	218
9.6	Low-Voltage Low-Power Operation	219
9.7	Reducing the Power Dissipation at the Architecture and Algorithm Levels	223

10. AMPLIFIER DESIGN FOR WIRELESS COMMUNICATION SYSTEMS	227
10.1 Introduction	227
10.2 Amplifier Design	228
10.3 Low Noise Amplifier	233
10.4 Automatic Gain Control Amplifiers	238
10.5 Power Amplifiers	243
11. PHASE LOCKED LOOPS	255
11.1 Introduction	255
11.2 Operation of the Phase Locked Loop	255
11.3 Phase Detectors	261
11.4 Frequency Dividers	269
11.5 Oscillator Design	273
12. FREQUENCY SYNTHESIZERS	285
12.1 Introduction	285
12.2 Frequency synthesizer (FS) parameters	286
12.3 Frequency Synthesizer Techniques	288
12.4 Analyzing phase noise in frequency synthesizers	299
12.5 Summary	303
Bibliography	306
Index	317

# List of Figures

1.1	Frequency assignments made to terrestrial and satellite FPLMTS (IMT-2000) systems by WARC-92.	3
1.2	Classification of wireless communication systems.	4
1.3	Electromagnetic spectrum of radio waves.	5
1.4	A generic transceiver.	6
1.5	A single-stage up conversion transmitter.	7
1.6	Direct conversion (homodyne) receiver.	8
1.7	A Two-stage superheterodyne receiver.	8
1.8	Allocated frequency spectrum for the AMPS system.	9
1.9	Frame structure for IS-136.	14
1.10	Slot format for a forward link digital traffic channel.	14
1.11	Slot format for a reverse link digital traffic channel.	15
1.12	Coding of a voice channel in IS-136.	15
1.13	Signal processing in the forward link transmitter of a CDMA system.	18
1.14	Signal processing in the reverse link transmitter of a CDMA system.	19
2.1	A wireless telecommunications system.	24
2.2	Generic block diagram of a digital transmitter.	25
2.3	Generic block diagram of a digital receiver.	26
2.4	Inter-symbol interference-free filters.	29
2.5	The transfer function of a raised cosine filter.	31
2.6	The impulse response of a raised cosine filter.	32
2.7	The impulse response of a realizable approximation of the impulse response of a raised cosine filter.	33
2.8	The input impulse sequence to a raised cosine filter having a truncated and delayed impulse response.	34
2.9	The output response of a raised cosine filter having a truncated and delayed impulse response.	35

2.10	The peak-to-peak jitter, $J_{pp}$ , versus the roll-off factor, $\alpha$ , of a raised cosine filter.	37
2.11	A rate-1/2 convolution encoder, with a constraint length of 3.	41
2.12	The finite state machine representation of a convolution encoder.	41
2.13	The trellis diagram of a convolution encoder.	42
2.14	Interleaving and De-interleaving.	42
2.15	Block Interleaving and de-interleaving.	43
2.16	Convolution Interleaver and De-interleaver.	44
2.17	Additive white Gaussian noise channel model.	46
2.18	Conditional probability density function at the output of the receiver's sampler.	48
2.19	Correlation receiver.	51
2.20	Wireless system satisfying the Nyquist criteria for no inter-symbol interference and the matched filter criteria for optimum bit error rate performance.	53
3.1	Cell shape.	57
3.2	Base station location relative to the cell boundary.	58
3.3	Co-channel interference in the cellular system.	58
3.4	Relative positioning of two cells allocated the same portion of the frequency spectrum in a cellular system.	59
3.5	Cellular system cell plan having a cluster size of seven cells.	60
3.6	Cell splitting in cellular systems.	63
3.7	Intersystem handoff.	66
3.8	Two ray propagation model: Direct line of sight (LOS) ray and ground reflected ray.	70
3.9	Illustration of the Doppler shift.	73
3.10	Small scale fading in a multi-path time-variant channel.	76
3.11	Rayleigh fading: Multi-wave reception.	79
3.12	Probability distribution function (pdf) and cumulative distribution function (CDF) for a Rayleigh distributed random variable.	81
3.13	Envelope variation of a band-limited Rayleigh distributed random process versus time.	82
3.14	Envelope probability density function in a Ricean fading environment.	84
3.15	A three rake finger receiver.	86
4.1	Amplitude shift keying (ASK) waveforms.	88
4.2	The power spectral density of an amplitude shift keying (ASK) modulated signal.	89

4.3	Bit error rate for coherent and non-coherent amplitude shift keying (ASK) demodulation.	90
4.4	Binary phase shift keying (BPSK) waveforms.	92
4.5	Differential binary phase shift keying (DBPSK) modulation and demodulation.	94
4.6	Differential binary phase shift keying (DBPSK) waveforms.	95
4.7	Signal-space diagram for quadrature phase shift keying (QPSK) modulation.	96
4.8	Quadrature phase shift keying (QPSK) modulator.	97
4.9	Quadrature phase shift keying (QPSK) waveforms.	98
4.10	Quadrature phase shift keying (QPSK) demodulator.	98
4.11	Offset quadrature phase shift keying (OQPSK) waveforms.	100
4.12	Differential quadrature phase shift keying (DQPSK) demodulator.	102
4.13	The space diagram of a $\pi/4$ -DQPSK system showing all allowable phase state transitions.	104
4.14	The carrier recovery circuit for a $\pi/4$ -DQPSK modulated signal.	105
4.15	Limiter discriminator detector for a $\pi/4$ -DQPSK modulated signal.	106
4.16	Differential detector for a $\pi/4$ -DQPSK modulated signal.	106
4.17	The bit error rate of a $\pi/4$ -DQPSK modulation scheme.	107
4.18	Minimum Shift Keying (MSK) waveforms.	108
4.19	Minimum Shift Keying (MSK) Modulator.	109
4.20	Minimum Shift Keying (MSK) Demodulator.	110
4.21	Pulse response of a Gaussian low-pass filter.	112
4.22	Gaussian Minimum Shift Keying (GMSK) Modulator.	112
4.23	Power spectrum density for digital modulation schemes.	113
4.24	Bit error rate for digital modulation schemes.	114
5.1	The operation of spread spectrum systems.	116
5.2	The direct sequence spread spectrum (DSSS) approach.	119
5.3	Waveforms of a direct sequence spread spectrum (DSSS) system.	119
5.4	The frequency hopping spread spectrum (FHSS) approach.	120
5.5	The time hopping spread spectrum (THSS) approach.	121
5.6	The Fibonacci Linear Feedback Shift Register.	122

5.7	The Galois Linear Feedback Shift Register.	123
5.8	A four stage linear feedback shift register used to generate PN sequences.	123
5.9	A direct sequence spread spectrum (DSSS) system.	127
5.10	Probability of false acquisition versus $E_b/N_o$ .	128
5.11	Early Late Tracker.	129
5.12	On-time, early and late correlator outputs of an early late tracker.	130
6.1	Noise figure in a one stage receiver.	134
6.2	The noise figure of a two stage receiver.	135
6.3	The noise figure of an n-stage receiver.	136
6.4	Receiver used in the calculation of the noise figure.	137
6.5	A two port non-linear network to illustrate non-linear distortion.	138
6.6	The frequency spectrum of the output of a third-order non-linear two-port network.	140
6.7	Intercept points for a third-order non-linear amplifier.	141
6.8	Intercept points of a two-stage non-linear receiver.	143
6.9	Intercept points of a $n$ -stage non-linear receiver.	143
6.10	Block diagram of a single-stage superheterodyne receiver.	145
6.11	Image frequency in a superheterodyne receiver.	146
6.12	A two-stage superheterodyne receiver.	147
6.13	Single side-band mixer.	148
6.14	Alternative single side-band mixer.	149
6.15	The homodyne (zero-IF) receiver.	150
6.16	Aliasing in a homodyne receiver.	151
6.17	The software radio architecture.	153
7.1	Block diagram of an analog-to-digital converter (ADC).	156
7.2	Block diagram of a digital-to-analog converter (DAC).	156
7.3	Block diagram of a digital communications system.	157
7.4	Non-linear transfer characteristic of an analog-to-digital converter.	159
7.5	Sampling models.	160
7.6	Ideal Sampling.	161
7.7	The spectrum of a sampled-signal sampled at less than the Nyquist rate.	162
7.8	Frequency spectrum of a pulse sampled signal.	164
7.9	Frequency spectrum of a band-pass signal.	165
7.10	Allowable band-pass sampling rates.	167
7.11	Staircase transfer function of a mid-tread quantizer.	168
7.12	Staircase transfer function of a mid-rise quantizer.	168

7.13	Quantization error versus input signal's magnitude for a mid-tread quantizer.	170
7.14	Probability density function of the quantization error.	170
7.15	Block diagram of an $n$ -bit flash analog-to-digital converter.	172
7.16	Transfer function of the amplifier used in the interpolative analog-to-digital converter.	174
7.17	Interpolative analog-to-digital converter with an interpolation factor of 2.	175
7.18	A three-bit interpolative analog-to-digital converter with an interpolation factor of three.	176
7.19	Relation between $(V_2^- - V_1^+)$ and $v_{in}$ for different values of the amplifier gain ( $A$ ), for the interpolative analog-to-digital converter of Figure 7.17.	178
7.20	A two-step analog-to-digital converter.	178
7.21	A two-stage pipelined analog-to-digital converter.	179
7.22	A two-step recycling analog-to-digital converter.	180
7.23	Block diagram of a subranging analog-to-digital converter.	181
7.24	Quantization intervals in a subranging analog-to-digital converter.	182
7.25	Block diagram of a folding analog-to-digital converter.	183
7.26	Block diagram of a folding circuit consisting of two amplifiers.	183
7.27	Block diagram of a folding analog-to-digital converter having a coarse analog-to-digital converter with 4 quantization intervals.	184
7.28	Folding analog-to-digital converter with a folding factor of 4.	185
7.29	Output waveform of a 4-bit successive approximation analog-to-digital converter.	187
7.30	Block diagram of a successive approximation analog-to-digital converter.	187
7.31	A Sigma-Delta Analog-to-Digital Converter.	188
7.32	Block diagram of a first-order Sigma-Delta modulator.	189
7.33	Frequency response of a first-order Sigma-Delta modulator.	191
7.34	Block diagram of a second-order Sigma-Delta modulator.	191
7.35	Block diagram of a second-order band-pass Sigma-Delta modulator.	193
7.36	The decimation filter.	193

7.37	The transfer function of a Sine decimator.	194
8.1	Partitioning of a wireless mobile terminal.	196
8.2	Block diagram of a Field Programmable Gate Array (FPGA).	202
8.3	The FPGA design flow.	203
8.4	The RF subsystem of a wireless communications system,	205
8.5	Single-stage heterodyne receiver.	206
8.6	Direct conversion receiver.	207
8.7	Heterodyne transmitter having a single IF stage.	208
8.8	Direct conversion transmitter.	208
8.9	Direct conversion transmitter using two local oscillators.	209
9.1	Four-input AND gate built using two-input AND gates.	215
9.2	A simplified system consisting of two building blocks and the interconnection busses.	216
9.3	The effect of reducing the supply voltage in CMOS circuits on delay and power dissipation.	220
9.4	Unbalanced pipeline example.	221
9.5	A two-datapath parallel system.	221
9.6	Combining parallelism with pipelining to balance pipe-stage delays.	222
9.7	Tree-Structured Vector Quantization.	225
10.1	Circuit diagrams of BJT and FET amplifiers	229
10.2	High-frequency hybrid pi-model of a bipolar junction transistor (BJT).	230
10.3	High-frequency hybrid pi-model of a field effect transistor (FET).	233
10.4	Noisy two-port network.	235
10.5	A gain-controlled differential amplifier.	238
10.6	A gain-controlled linearized differential amplifier.	239
10.7	A gain-controlled amplifier.	242
10.8	Class A power amplifier.	245
10.9	The output waveform of a class A power amplifier.	245
10.10	Class A power amplifier.	246
10.11	Load line of a class A power amplifier.	248
10.12	Class B power amplifier.	249
10.13	Transfer characteristics of a class B power amplifier.	250
10.14	Class C Power Amplifier.	252
10.15	Waveforms of the collector current and output voltage for a class C power amplifier.	253
10.16	Class F power Amplifier.	253
10.17	Collector current and voltage waveforms for a class F power amplifier.	254

11.1	Block diagram of a phase locked loop (PLL).	256
11.2	Block diagram of a phase locked loop (PLL) in the frequency-domain.	258
11.3	Step response of a phase locked loop.	260
11.4	Loop filter output versus phase error at the input of a multiplier phase detector.	263
11.5	Input and output waveforms to an XOR digital phase detector.	264
11.6	Output voltage of the loop filter versus the phase shift between the two inputs to an XOR digital phase detector.	265
11.7	Block diagram of a digital phase detector using edge-triggered flip flops.	265
11.8	State machine representation of the digital phase detector of Figure 11.7.	265
11.9	Transfer function of the digital phase detector of Figure 11.7.	266
11.10	Block diagram of an edge-triggered phase detector.	267
11.11	Timing diagram for the edge-triggered phase detector of Figure 11.10, having a $180^\circ$ phase shift between its inputs.	268
11.12	Timing diagram for the edge triggered phase detector of Figure 11.10, having higher frequency on the reference input.	269
11.13	Timing diagram for the edge triggered phase detector of Figure 11.10, having higher frequency on the oscillator input.	270
11.14	Toggle flip flop.	270
11.15	$n$ -stage ripple counter.	271
11.16	A synchronous divide-by-six counter.	271
11.17	Timing diagram for a divide-by-six counter.	272
11.18	A programmable comparator.	273
11.19	Block diagram of a positive feedback oscillator	274
11.20	Circuit diagram of a phase shift oscillator	275
11.21	Generalized circuit diagram of the Colpitts/Hartley Oscillator.	275
11.22	Small signal model for active devices.	276
11.23	Generalized equivalent circuit for the Colpitts/Hartley Oscillator.	276
11.24	Circuit diagram of a Colpitts Oscillator.	279
11.25	Circuit diagram of a Hartley Oscillator.	279
11.26	Equivalent circuit of the Colpitts/Hartley oscillator illustrating the oscillator's negative resistance.	280

11.27	Circuit diagram of the Clapp-Gouriet Oscillator.	281
11.28	Crystal oscillator.	282
11.29	Crystal oscillator reactance versus frequency.	283
11.30	Circuit diagram of the Pierce Oscillator	283
12.1	General frequency synthesizer structure.	285
12.2	Frequency representation of a sine wave (a) in ideal form and (b) corrupted by phase noise.	287
12.3	Typical output spectrum of a frequency synthesizer.	288
12.4	Fractional-N PLL architecture.	289
12.5	Block diagram of a generic fractional-N frequency divider.	291
12.6	Divide-by-4/5 prescaler.	291
12.7	Typical values stored in an accumulator of a pulse swallower counter.	292
12.8	Frequency spectrum of spurs in a fractional-N divider [1].	293
12.9	Block diagram of API used in a fractional-N PLL synthesizer.	294
12.10	Digital Sigma-Delta modulator.	294
12.11	Conventional first-order Sigma-Delta modulator.	295
12.12	Effect of Sigma-Delta modulator on spurs of fractional-N synthesizers [2].	296
12.13	Direct digital frequency synthesizer architecture.	297
12.14	Phase noise in a closed-loop PLL system.	302

# List of Tables

1.1	Spectrum allocation for mobile communications systems.	6
1.2	Comparison of analog and digital communication techniques.	8
1.3	Vocoder output rates.	17
2.1	Zero crossing time variations of a raised cosine filter.	36
2.2	Code-words for a (7,4) Systematic Block Code.	40
3.1	Path loss exponent for different wireless environments.	72
4.1	Differential encoding for quadrature phase shift keying.	101
5.1	Number of one and zero runs of length $n$ of a PN sequence generated using a 5-stage linear feedback shift register.	123
5.2	Number of primitive polynomials of degree $N$ , $N = 2 - 6$ .	124
5.3	The output sequence of the linear feedback shift register of Figure 5.8.	125
6.1	Power gain, noise figure, effective noise temperature for each stage of the receiver shown in Figure 6.4.	138
7.1	The thermometer code of a binary coded signal.	172
7.2	Latch outputs for an interpolative analog-to-digital converter with an interpolation factor of 2.	175
7.3	Dynamic range versus OSR for first, second and third order Sigma-Delta modulators.	192
9.1	Vector quantization: Computation and memory requirements.	224
12.1	Conventional versus Fractional-N Frequency Dividers.	290

# Preface

“Wireless is coming” was the message received by VLSI designers in the early 1990’s. They believed it. But they never imagined that the wireless wave would be coming with such intensity and speed. Today one of the most challenging areas for VLSI designers is VLSI circuit and system design for wireless applications.

New generation of wireless systems, which includes multimedia, put severe constraints on performance, cost, size, power and energy. The challenge is immense and the need for new generation of VLSI designers, who are fluent in wireless communication and are masters of mixed signal design, is great.

No single text or reference book contains the necessary material to educate such needed new generation of VLSI designers. There are gaps. Excellent books exist on communication theory and systems, including wireless applications and others treat well basic digital, analog and mixed signal VLSI design. We feel that this book is the first of its kind to fill that gap.

In the first half of this book we offer the reader (the VLSI designer) enough material to understand wireless communication systems. We start with a historical account. And then we present an overview of wireless communication systems. This is followed by detailed treatment of related topics; the mobile radio, digital modulation and schemes, spread spectrum and receiver architectures.

The second half of the book deals with VLSI design issues related to mixed-signal design. These include analog-to-digital conversion, transceiver design, digital low-power techniques, amplifier design, phase locked loops and frequency synthesizers.

The book can be used in a senior undergraduate or graduate course on VLSI Design For Wireless Applications and can be also used by practicing engineers and managers working in this area. Basic background in both communication systems and VLSI design are assumed.

We are aware of the hazards of trying , in a 300-plus page book, to bridge the gap between two engineering disciplines. Trading depth and breadth is only one such hazard. But our conviction for the need of such a book motivated us to write it. Moreover, the experience of writing it was enjoyable and we hope the book proves useful to our readers.

Emad Farag

Mohamed Elmasry

Waterloo, Ontario, Canada

## **Acknowledgments**

We first acknowledge the blessings of God Almighty in our lives. The contributions of many people to this book, through discussions, is very much appreciated. The partial support received from Canadian Research granting agencies, and from Canadian, U.S. and Japanese industry and the University of Waterloo is acknowledged.

Lucent Technology has provided to one of us (Farag) an excellent environment to complete this book. The VLSI Research Group at the University of Waterloo has been instrumental during the different phases of writing and we thank the faculty, staff and students.

Finally, our families offered the necessary support to complete this book and we thank them dearly.

## List of Acronyms

3G:	Third Generation
ACI:	Adjacent Channel Interference
ACTS:	Advanced Communications Technologies and Services
ADC:	Analog-to-Digital Converter
AGC:	Automatic Gain Control
AM:	Amplitude Modulation
AMPS:	Advanced Mobile Phone Service
ARIB:	Association of Radio Industries and Businesses
ASIC:	Application Specific Integrated Circuit
ASK:	Amplitude Shift Keying
AWGN:	Additive White Gaussian Noise
BCCH:	Broadcast Control Channel
BER:	Bit Error Rate
BiCMOS:	Bipolar Complementary Metal Oxide Semiconductor
BJT:	Bipolar Junction Transistor
BPSK:	Binary Phase Shift Keying
CDF:	Cumulative Distribution Function
CDMA:	Code Division Multiple Access
CDPD:	Cellular Digital Packet Data
CDVCC:	Code Digital Verification Color Code
CELP:	Code Excited Linear Prediction
CEPT:	Conference of European Posts and Telecommunications
CMOS:	Complementary Metal Oxide Semiconductor
CRC:	Cyclic Redundancy Check
CSMA:	Carrier Sense Multiple Access
DAC:	Digital-to-Analog Converter
DBPSK:	Differential Binary Phase Shift Keying
DCCH:	Dedicated Control Channel
DDFS:	Direct Digital Frequency Synthesis

DECT:	Digital European Cordless Telephone
DDI:	Daini-Denden, Inc.
DQPSK:	Differential Quadrature Phase Shift Keying
DSSS:	Direct Sequence Spread Spectrum
DSP:	Digital Signal Processing/Processor
DTC:	Dedicated Traffic Channel
EEPROM:	Electrical Erasable Programmable Read Only Memory
EIRP:	Effective Isotropic Radiated Power
EPROM:	Erasable Programmable Read Only Memory
ETACS:	Extended Total Access Communication System
ETSI:	European Telecommunications Standards Institute
FACCH:	Fast Associated Control Channel
FCC:	Federal Communications Commission
FCW:	Frequency Control Word
FDD:	Frequency Division Duplexing
FDMA:	Frequency Division Multiple Access
FEC:	Forward Error Correction
FET:	Field Effect Transistor
FHSS:	Frequency Hopping Spread Spectrum
FM:	Frequency Modulation
FPLMTS:	Future Public Land Mobile Telecommunications System
FS:	Frequency Synthesizer
FSK:	Frequency Shift Keying
FSVQ:	Full Search Vector Quantization
GaAs:	Gallium Arsinide
GHz:	Giga Hertz
GMSK:	Gaussian Minimum Shift Keying
GSM:	Global System for Mobile communications previously known as Group Speciale Mobile
HDL:	Hardware Description Language
IF:	Intermediate Frequency
IMPS:	Improved Mobile Phone System
IMT-2000:	International Mobile Telecommunications 2000
IS:	Interim Standard
ISI:	Intersymbol Interference
ITU:	International Telecommunications Union
JFET:	Junction Field Effect Transistor
JTACS:	Japanese Total Access Communication System
JTC:	Japanese Cordless Telephone
kbps:	Kilo bits per second
Kcps:	Kilo chips per second
LAN:	Local Area Network

LNA:	Low Noise Amplifier
LOS:	Line of Sight
Mbps:	Mega bits per second
Mcps:	Mega chips per second
MHz:	Mega Hertz
MIPS:	Million Instructions Per Second
MMIC:	Microwave Monolithic Integrated Circuit
MOSFET:	Metal Oxide Semiconductor Field Effect Transistor
MSC:	Mobile Switching Center
MSK:	Minimum Shift Keying
MTS:	Mobile Telephone System
NAMPS:	Narrow-band Advanced Mobile Phone Service
NF:	Noise Figure
NMOS:	N-type Metal Oxide Semiconductor
NMT:	Nordic Mobile Telephone
NTACS:	Narrow-band Total Access Communication System
NTT:	Nippon Telephone and Telegraph
OOK:	On Off Keying
OQPSK:	Offset Quadrature Phase Shift Keying
PCS:	Personal Communications System
PDC:	Personal Digital Cellular
pdf:	Probability Density Function
PHS:	Personal Handy (phone) System
PLL:	Phase Locked Loop
PN:	Pseudorandom Noise
ppm:	Part-Per-Million
PSD:	Power Spectral Density
PSK:	Phase Shift Keying
PSTN:	Public Switched Telephone Network
QAM:	Quadrature Amplitude Modulation
QPSK:	Quadrature Phase Shift Keying
RACE:	Research into Advanced Communications in Europe
RAM:	Random Access Memory
RF:	Radio Frequency
ROM:	Read Only Memory
QAM:	Quadrature Amplitude Modulation
QPSK:	Quadrature Phase Shift Keying
RACH:	Random Access Channel
RAM:	Random Access Memory
RSSI:	Received Signal Strength Indicator
SACCH:	Slow Associated Control Channel
SCF:	Shared Channel Feedback

SiGe:	Silicon Germanium
SMS:	Short Message Service
SNR:	Signal-to-Noise Ratio
SOI:	Silicon-On-Insulator
SPACH:	SMS Point-to-Point and Access Response Channel
TAGS:	Total Access Communication System
TDD:	Time Division Duplexing
TDMA:	Time Division Multiple Access
THSS:	Time Hopping Spread Spectrum
TIA:	Telecommunications Industry Association
TSVQ:	Tree Structured Vector Quantization
UE:	User Equipment
UMTS:	Universal Mobile Telecommunications System
VCO:	Voltage Controlled Oscillator
VHDL:	VHSIC Hardware Description Language
VHSIC:	Very High Speed Integrated Circuit
VLSI:	Very Large Scale Integration
VSELP:	Vector Sum Excited Linear Prediction
W-CDMA:	Wideband Code Division Multiple Access
WARC:	World Administrative Radio Conference

# Chapter 1

## INTRODUCTION

### 1.1 THE HISTORY OF WIRELESS COMMUNICATIONS

The first use of electricity as a means of conveying information dates back to the first half of the nineteenth century. In 1837, Samuel Morse invented the telegraph. The first public telegram was sent in 1844. A few decades later in 1876, Alexander Graham Bell was able to convey human voice through electrical signals. This invention was called the telephone.

The use of electromagnetic waves as a means of conveying information between two untethered terminals started over a century ago. In 1887, Heinrich Hertz, using high frequency currents, was able to illustrate the propagation of an electromagnetic wave from a transmitter to a receiver. In 1894, Gugliemo Marconi began his research on wireless communication systems which eventually led to the first patent on wireless communications to be issued to him in 1896.

The first two-way mobile communications systems were used by the police departments during the 30's and 40's. Initially, these systems used amplitude modulation. However, by 1940 most of them converted to frequency modulation because of its higher resilience to propagation impairments. During World War II mobile communications was used extensively by the military.

The first public mobile phone system was deployed in the United States just after the war, in 1946, it was called the Mobile Telephone System (MTS). This system was a simplex system, which required the intervention of an operator to place calls. In 1969, an improved mobile system was introduced, known as the Improved Mobile Phone System (IMPS). This system, which was also deployed in the United States, had full duplex operation and automatic switching. It operated at a frequency of 450 MHz.

In Japan, Nippon Telegraph and Telephone (NTT) started research into mobile systems in 1953, and by 1967 the development was complete. The first mobile system was non-cellular, but it was fully automatic, and operated at a frequency band around 450 MHz. This system didn't go into commercial use because of the lack of available spectrum in the 450 MHz band [3].

The American IMPS system quickly ran into congestion problems because it used one large zone per city. In 1983, a more advanced mobile communications system called the Advanced Mobile Phone Service (AMPS) succeeded it. AMPS employs cellular concepts to increase the spectral efficiency of the mobile system. By 1996, there were about 35 million AMPS subscribers in the United States.

Other analog cellular systems were introduced during the late 70's and early 80's in different parts of the world. The first cellular system to be deployed was in 1979, when Nippon Telegraph and Telephone (NTT) launched its cellular phone service in Japan. In 1989, a more advanced narrow-band high capacity system was introduced. In Europe, several analog cellular standards were introduced, such as the Nordic Mobile Telephone in the Nordic countries, the Total Access Communication System (TAGS) in the United Kingdom, RC-2000 in France, C-net in Germany and I-450 in Italy. The analog cellular systems later became known as the First Generation cellular systems.

The capacity of the analog cellular systems couldn't meet the rapid increase in demand. Furthermore, because many different analog cellular systems existed in Europe, there was a lack in the ability of roaming from one country to the other. This prompted the standardization of a Pan-European digital cellular system. This system is called GSM (Global System for Mobile Communications), it was standardized by the Conference of European Posts and Telecommunications (CEPT) in 1988. By November 1996, the number of GSM subscribers (19 million) exceeded the number of analog cellular subscribers (13 million) in Europe. Currently there are about 150 million GSM subscribers world wide.

In the United States, the Telecommunications Industry Association (TIA) issued two interim standards for Second Generation digital cellular systems. IS-54 was issued in 1989, it was later reissued as IS-136, this is an interim standard for a time division multiple access (TDMA) based digital cellular system. IS-95 was issued in 1993. This is an interim standard for a code division multiple access (CDMA) based digital cellular system. The Japanese digital cellular system, which is called Personal Digital Cellular was standardized in 1992 by the Research and Development Center for Radio Systems (CRC) which was later renamed to the Association of Radio Industries and Businesses (ARIB). By March 1996, there were 10 million analog and digital cellular subscribers in Japan.

Analog and then digital cellular systems were designed with the goal of carrying voice, fax and low bit rate data. The bit rate carried per channel ranges from 48.6 kbps, for 3 users in the IS-136 standard, to 270.833 kbps, for 8 users in GSM. These low bit rates are insufficient to carry multimedia information. This fact coupled with the unprecedented growth of demand for mobile phones especially in Europe and Japan, for example in Europe the number of digital cellular subscribers doubled in about 10 months in 1996, and the pressure to integrate fixed and mobile networks, points to the urgent need for a flexible and cost effective Third Generation (3G) mobile communications system. This Third Generation system should integrate the capabilities of cellular, cordless and wireless LAN systems by providing a high bit rate wireless link capable of serving terminals with high mobility. Third Generation mobile systems are expected to have a quality of service similar to that of fixed networks, and to operate in different environments for high speed and low speed mobile terminals and to operate indoors as well as outdoors. Third generation mobile communication systems are expected to use wideband Code Division Multiple Access (W-CDMA) as their radio access technology [4].

The ITU-R, formerly known as CCIR, started the standardization of the Future Public Land Mobile Telecommunications System (FPLMTS) which was later renamed IMT-2000 (International Mobile Telecommunications 2000) in 1995. The 1992 World Administrative Radio Conference (WARC-92) of the International Telecommunications Union (ITU) allocated, on a world wide basis, 230 MHz in the 2 GHz band (1885-2025 MHz and 2110-2200 MHz), to Third Generation wireless communications systems. This frequency allocation includes both satellite and terrestrial links, as shown in Figure 1.1.

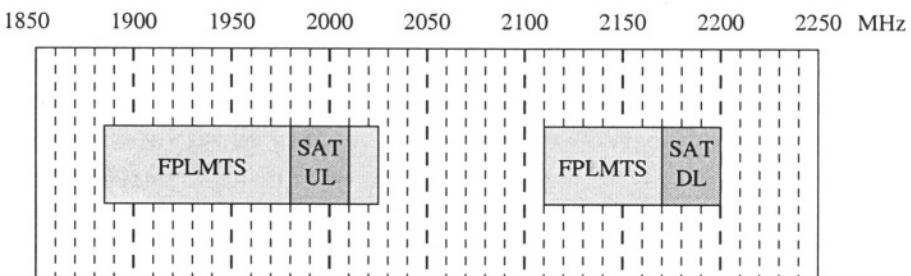


Figure 1.1. Frequency assignments made to terrestrial and satellite FPLMTS (IMT-2000) systems by WARC-92.

In Europe the Third Generation wireless communications system is being developed under the name, Universal Mobile Telecommunications System (UMTS). The standardization process for UMTS started in 1990 by the European Telecommunications Standards Institute (ETSI), and has been the subject of extensive research which was carried out mainly in the context

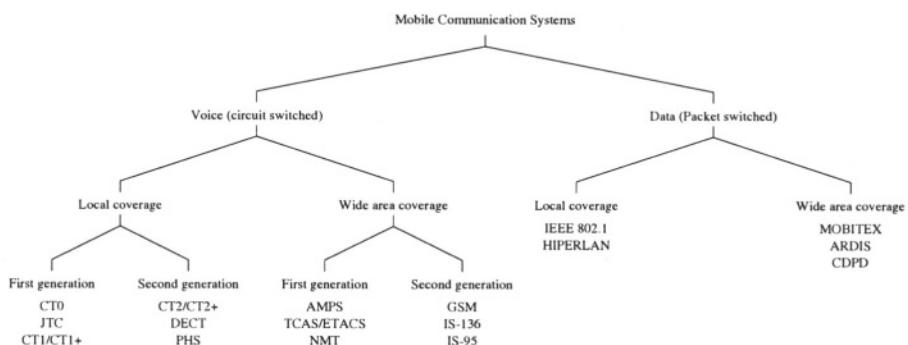
of the European Community research programs, such as RACE (Research into Advanced Communications in Europe), and then ACTS (Advanced Communications Technologies and Services).

Early next century is the target set for the deployment of Third Generation systems in many areas of the world. Japan leads the way by targeting the year 2000 for the initial deployment of Third Generation systems. A second phase, with much higher rates (10 Mbps versus 2 Mbps) is expected to be deployed by 2005. Europe is targeting 2002 for the implementation of basic UMTS service, full UMTS capabilities should be implemented by the year 2005. However, in North America, the huge First Generation legacy, and the recent investments in the Second Generation systems, has led to a relaxed approach, targeting 2005 for the deployment of Third Generation systems. However, incremental enhancements are expected to Second Generation systems.

## 1.2 WIRELESS COMMUNICATIONS BASICS

### 1.2.1 Mobile Wireless Communication Systems

The increase in consumer demand for wireless systems that enable consumers to communicate in any place, by any means of information has led to the emergence of many portable wireless communication products. Many wireless services are being accepted by the public such as cellular and personal communications systems, cordless systems, paging, and wireless data transmission systems, which include high-speed wireless local area networks, and wide area public data networks. Figure 1.2 shows the classification of wireless communications systems into data and voice wireless systems as well as local and wide area coverage systems.



*Figure 1.2. Classification of wireless communication systems.*

## 1.2.2 Spectrum Allocations

Radio waves are part of the electromagnetic spectrum with a frequency range extending from a few KHz's to 100's of GHz, which corresponds to a wavelength in the range  $3 \times 10^5$  m to 3 mm. Radio waves are used as a means of conveying information between the transmitter and the receiver of a wireless communications system. Figure 1.3 illustrates the electromagnetic spectrum used for radio transmissions. Table 1.1 shows the frequency spectrum allocated to different mobile communication systems.

Frequency Band	Application	Frequency	Wavelength
ELF Extremely Low Frequency	Submarine	30 Hz	10000 Km
ULF Ultra Low Frequency		300 Hz	1000 Km
VLF Very Low Frequency	Navigation	3 KHz	100 Km
LF Low Frequency	Maritime mobile LW broadcast	30 KHz	10 Km
MF Medium Frequency	MW broadcast	300 KHz	1 Km
HF High Frequency	SW broadcast	3 MHz	100 m
VHF Very High Frequency	FM broadcast	30 MHz	10 m
UHF Ultra High Frequency	TV, mobile LOS microwave	300 MHz	1 m
SHF Super High Frequency	LOS microwave Satellite	3 GHz	10 cm
EHF Extremely High Frequency		30 GHz	1 cm
		300 GHz	1 mm

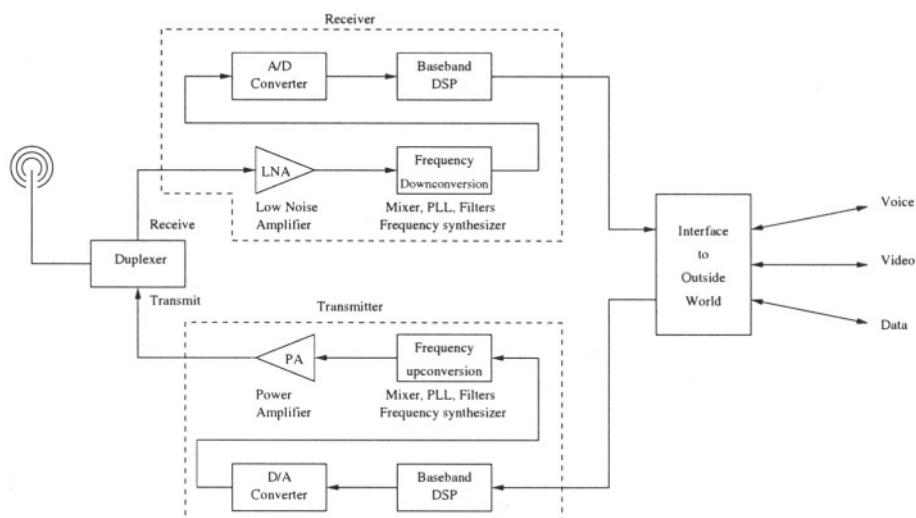
Figure 1.3. Electromagnetic spectrum of radio waves.

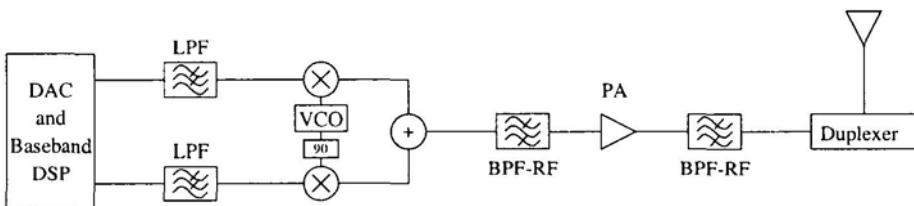
*Table 1.1.* Spectrum allocation for mobile communications systems.

System	Type	Frequency Range (MHz)	
		Up Link	Down Link
AMPS	Analog Cellular	824 - 849	869 - 894
TACS	Analog Cellular	890 - 915	935 - 960
NMT-450	Analog Cellular	453 - 458	463 - 468
NMT-900	Analog Cellular	890 - 915	935 - 960
IS-136	Digital Cellular	824 - 849	869 - 894
IS-95	Digital Cellular	824 - 849	869 - 894
GSM	Digital Cellular	890 - 915	935 - 960
PDC	Digital Cellular	940 - 956	810 - 826
PDC	Digital Cellular	1477 - 1501	1429 - 1453
IMT-2000	Third Generation	1920 - 1980	2110 - 2170
DECT	Digital Cordless	1880 - 1990	1880 - 1990
PHS	Digital Cordless	1895 - 1907	1895 - 1907
DCS 1800	Digital Cordless	1710 - 1785	1805 - 1880
CDPD	Wireless Data	824 - 849	869 - 894
Mobitex	Wireless Data	896 - 902	935 - 941
Ardis	Wireless Data	806 - 824	851 - 869

### 1.2.3 A Generic Transceiver

Figure 1.4 shows the block diagram of a generic transceiver. The upper blocks represent the receive part of the transceiver, while the lower blocks represent the transmit part of the transceiver.

*Figure 1.4.* A generic transceiver.



**Figure 1.5.** A single-stage up conversion transmitter.

In the transmitter chain, the information signal, which can be speech, video or data, is digitized and organized into frames. This data passes through several stages of coding and digital signal processing to transform it into a format suitable for transmission across a noisy fading wireless channel. The digital signal processing stages include error correction and detection, interleaving, time division multiplexing and/or spreading. The signal is then modulated onto a carrier and up-converted to the desired transmit frequency.

In the transmitter the baseband signal is up-converted to the RF carrier frequency. This can be done using a single-stage quadrature modulator [5]. This is shown in Figure 1.5. A filter is needed after the power amplifier. This filter removes any out of band frequency components due to the non-linearity of the power amplifier or spurious frequencies from the oscillator. A transmitter can also have multi-stage mixing.

In the receive chain, the signal amplified by the low noise amplifier (LNA) is a very weak signal, which can be as low as -100 dBm<sup>1</sup> or even lower, any noise added by the LNA can corrupt the received signal significantly and subsequently degrades the overall system performance. Hence, an important consideration in the design of the LNA is the minimization of the amount of noise it generates. The noise generated by the LNA is described by its noise figure.

The received signal has a center frequency of a few 100's of MHz, or even a few GHz. The frequency down conversion stage locks to the frequency of the received signal and down converts it to baseband, as in a homodyne receiver (shown in Figure 1.6), or to a lower intermediate frequency (IF), as in a superheterodyne receiver (shown in Figure 1.7). This down conversion is done by mixing the received signal with a sinusoidal carrier derived from the center frequency of the received signal, and generated by a frequency synthesizer.

The baseband (or IF) signal is digitized and the baseband processing is done digitally by a digital signal processor (DSP) and/or an application specific integrated circuit (ASIC). Digital baseband processing in the receiver involves

---

<sup>1</sup> Power in dBm =  $10 \log \left( \frac{\text{Power}}{1 \text{ mW}} \right)$ .

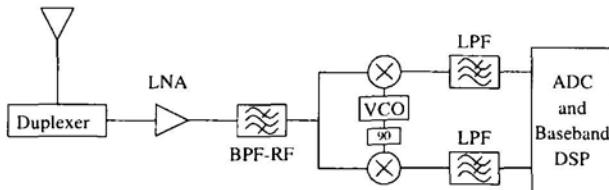


Figure 1.6. Direct conversion (homodyne) receiver.

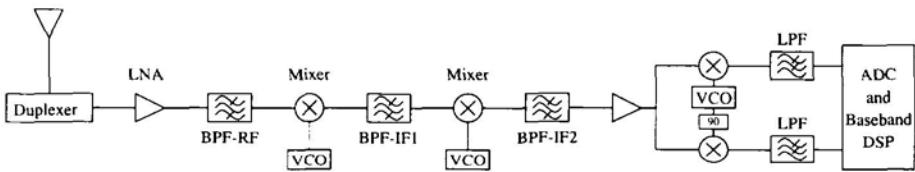


Figure 1.7. A Two-stage superheterodyne receiver.

Table 1.2. Comparison of analog and digital communication techniques.

Criterion	Analog Communication Techniques	Digital Communication Techniques
Technology	Usually Discrete/Hybrid	Usually Monolithic
Modulation	FM	QPSK, GMSK ...
Access	FDMA	TDMA, CDMA
Noise resilience	Low	High
Voice/Video Compression	Not possible	Possible
Security	Low	High
Capacity (Spectral efficiency)	Low	High

code despreading, time demultiplexing, channel decoding, source decoding, etc. Eventually, the signal is converted back into its original form, an analog speech or video signal, or a digital data signal.

In the last few years, there has been a shift from analog to digital communication techniques. This shift led to enhanced performance and lower cost. Table 1.2 compares analog communications techniques to digital ones. Digital systems are more flexible and dynamic in operation than analog systems, for example digital systems allow the varying of the channel bit rate to better suit the needs of the transmitted signal. In the future, this digitization trend is expected to continue moving into the front-end of the transceiver, and eventually leading to a true software radio.

## 1.3 WIRELESS COMMUNICATIONS STANDARDS

### 1.3.1 AMPS

The Advanced Mobile Phone Service (AMPS) is an analog cellular system that is dominantly used in North America. In the late 70's the FCC (Federal Communications Commission) issued licenses in Washington and Chicago, to evaluate the AMPS specification. During the 80's the AMPS system started its penetration into the North American markets.

The frequencies allocated to the AMPS system are between 824 MHz to 849 MHz for the reverse channel, and 869 MHz to 894 MHz for the forward channel. The transmit-receive bands are separated by 45 MHz. Each channel has a bandwidth of 30 KHz, thus the total number of channels is 832. These channels are divided equally between two operators (A and B), as shown in Figure 1.8. The A frequencies are for a non-wirelined operator, while the B frequencies are for a wirelined operator. AMPS uses FM modulation.

# of Channels	$A_2$ 33 Channels	A 333 Channels	B 333 Channels	$A_1$ 50 Channels	$B_1$ 833 Channels
Reverse Link	824.04 825.03		835.02		845.01 846.51
Forward Link	869.04 870.03		880.02		890.01 891.51

849 MHz      894 MHz

Figure 1.8. Allocated frequency spectrum for the AMPS system.

Initially, the FCC allocated 20 MHz for each of the forward and reverse links, which corresponds to 666 channels. At a later date, 5 MHz were added, giving 166 more channels. This is shown in Figure 1.8, the channels in bands  $A_1$ ,  $A_2$  and  $B_1$  are the channels that were added at a later time.

### 1.3.2 TACS and ETACS

Total Access Communication System (TASC) is an analog cellular system which was initially deployed in the United Kingdom and then Ireland in 1985. In 1990 it was deployed in other European countries: Austria, Italy, Malta and Spain [6].

In the UK the frequencies allocated to the TACS system are between 890 MHz to 915 MHz for the reverse link and between 935 MHz to 960 MHz for the forward link. Each channel has a bandwidth of 25 KHz. A telephone conversation requires two channels, one in each band. Thus each band has a 1000 channels, which were divided among two cellular operators: Vodafone and Cellnet, each was allocated 300 channels. The remaining 400 channels were reserved for the digital GSM system.

Of the 300 channels assigned to each operator 277 channels are used for speech, 21 channels are control channels and 2 channels are used as guard channels between the two cellular operators.

Due to the increase in demand, the British government allocated additional frequencies in the bands between 872 MHz to 888 MHz and 917 MHz to 933 MHz. The additional channels are known as ETASC (Extended Total Access Communications System). The additional bands gave a further 320 channels to each cellular operator [6]. The 320 channels assigned for each operator are speech channels. The same control channels are used for both TAGS and ETACS.

### **1.3.3 Nordic Mobile Telephone**

In the late 70's telecommunication officials from Denmark, Finland, Norway and Sweden, agreed on a common analog cellular system, which became to be known as the Nordic Mobile Telephone (NMT). The first NMT system was deployed in 1981. The first generation of NMT systems operated at 450 MHz, and is known as NMT-450. The reverse link frequency band extends from 453 MHz to 458 MHz, while the forward link frequency band extends from 463 MHz to 468 MHz. 10 MHz separates the receive/transmit channels. With a channel spacing of 25 KHz, the NMT-450 can support 200 channels. The coverage area of each base station is 15-40 Km.

A second generation of the NMT system was deployed in 1986. It operates at a frequency around 900 MHz and is hence known as NMT-900. The reverse link frequency band extends from 890 MHz to 915 MHz, while the forward link frequency band extends from 935 MHz to 960 MHz. The receive/transmit channels are separated by 45 MHz. With a channel spacing of 25 KHz, the NMT-900 can support 1000 channels. The NMT-900 system can also operate in an interleaved mode where the channels are interleaved in the frequency domain. In this case, with a channel spacing of 12.5 KHz, the system capacity becomes 1999 channels. The coverage area of each base station, in the NMT-900 system, is 2-20 Km.

The NMT system has been adopted by many countries worldwide, in addition to the Nordic countries, such as Romania, Poland, Turkey, Malaysia, Morocco, etc.

### **1.3.4 Analog Cellular Standards in Japan**

The first commercial cellular system in the world was the Japanese MCS-L1 (first generation Mobile Control Station) [3]. It was introduced in Japan in 1979. The forward link frequency band extends from 870 MHz to 885 MHz. The system can accommodate 600 channels with a channel spacing of 25 KHz.

Phase modulation, with a maximum frequency deviation of 5 KHz, is employed in the modulation of the voice signal.

The MCS-L1 quickly ran into congestion problems, as a result NTT started research and development on a higher capacity second generation analog system known as MCS-L2. In 1986, the Japanese Ministry of Posts and Telecommunications approved this system, and it was deployed in 1988. The forward and reverse link frequency bands are the same as those employed in MCS-L1. MCS-L2 has a channel spacing of 12.5 KHz, which can be further reduced to 6.25 KHz, when the channels are interleaved in the frequency domain. This increases the number of RF channels the system can accommodate up to 2400 channels. Phase modulation, with a maximum frequency deviation of 2.5 KHz, is employed in the modulation of the voice signal.

In 1989, a subsidiary of Daini-Denden, Inc. (DDI) launched its cellular service. This cellular system is based on the British TACS system, except that it uses different frequency bands. The forward link frequency band extends from 843 MHz to 846 MHz and from 860 MHz to 870 MHz. The reverse link frequency band extends from 898 MHz to 901 MHz and from 915 MHz to 925 MHz. The channel spacing is 50 KHz (25 KHz with interleaving). Hence, the total number of channels JTACS can support is 520 channels with interleaving. Phase modulation, with a maximum frequency deviation of 9.5 KHz, is employed in the modulation of the voice signal.

In 1991, a subsidiary of DDI and Nippon Idou Tsushin Crop. (IDO) launched the Motorola's NTACS system. This system is based on the NAMPS system. The NTACS system uses the same frequency bands as the JTACS system. The channel spacing is halved, 25 KHz for non-interleaved channels, and 12.5 KHz for interleaved channels. Hence, the capacity is doubled. NTACS employs phase modulation with a maximum frequency deviation of 5 KHz.

### **1.3.5 Global System for Mobile Communications**

The Global System for Mobile communications (GSM) was introduced in Europe in the early 90's at a time when first generation analog cellular systems reached their capacity limits. GSM is a pan-European digital cellular system, that allows roaming from one country to the other across Europe. This was not the case for first generation systems, where each country had its own analog cellular system, that was incompatible with that of other countries.

GSM is a second generation cellular system that employs digital radio technology. Second generation systems have several advantages over their predecessors, the first generation cellular systems which employed analog modulation techniques:

1. Higher spectral efficiency.

2. Greater compatibility. When the first generation systems were developed every country had a system that was incompatible with that of other countries. There is only one or two digital cellular systems in each part of the world. For example in Europe the digital cellular system is GSM.
3. Speech privacy.
4. New services which can be combined with the transmitted digital voice signal.
5. More robust to wireless channel impairments, when using appropriate channel coding.

GSM uses a carrier frequency separation of 200 KHz with a digital transmission bit rate of 270 Kbits/sec. GSM can accommodate 8 full rate channels per carrier at 16 Kbits/sec, or 16 half rate channels per carrier at 32 Kbits/sec. The mobile terminal transmits in the band (reverse channel) 890-915 MHz. While, the base station transmits in the band (forward channel) 925-960 MHz. Thus, the duplex channel separation is 45 MHz.

GSM uses slow frequency hopping spread spectrum (217 hops/sec) to combat Rayleigh fading when the mobile terminal is stationary or moving very slowly and happens to be in a spectral null.

The coding of speech is done using linear predictive coding at a rate of 13 Kbits/sec or 5.6 Kbits/sec. Voice activity detection is used to prevent transmission during silence. Power level control is also used to reduce the transmission power when the mobile terminal is close to the base station. The GSM cellular system is designed to handle services such as telephony, and data transmission up to 9.6 Kbits/sec, as well as short message service (SMS) [7] of up to 180 characters that can be transmitted over the signaling channel.

### **1.3.6 IS-136 North America TDMA Standard**

The IS-136 standard is a dual mode cellular standard that supports both analog and digital communication modes. The analog mode of the IS-136 standard is based on the AMPS standard. While, the digital mode of the standard employs time division multiple access (TDMA). IS-136 is an extension of the IS-54 standard, it differs from IS-54 in that it defines digital control channels in addition to digital traffic channels.

The frequency assignments for the forward link and the reverse link in IS-136 are identical to those of the AMPS standard. The forward link is allocated 25 MHz in the frequency band 869 MHz to 894 MHz. While, the reverse link is allocated 25 MHz in the frequency band 824 MHz to 849 MHz. Each TDMA channel occupies a bandwidth of 30 KHz. This bandwidth is shared by 3 or 6 users. The modulation scheme used in IS-136 is  $\pi/4$ -DQPSK.

A square-root raised cosine filter with  $\alpha = 0.35$  is used in both the transmitter and the receiver.

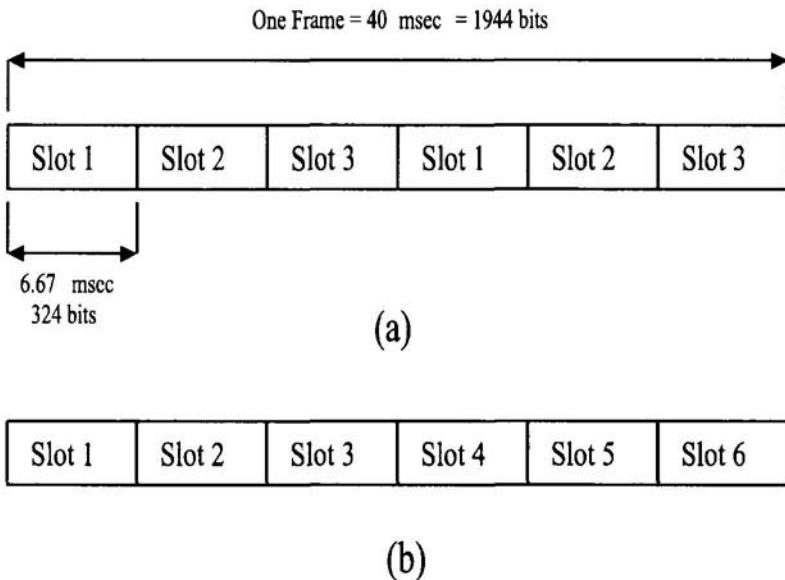
There are two types of logical channels specified in the IS-136 standard:

1. Dedicated Control Channels (DCCH). These are used to convey control information and short user-data messages between the base station and the mobile terminal. The following control channels are specified in IS-136:
  - (a) Shared Channel Feedback (SCF). This is a forward link control channel that carries status information (e.g. busy, idle).
  - (b) Broadcast Control Channel (BCCH). This is a forward link control channel that carries generic system information.
  - (c) SMS Point-to-Point, Paging and Access Response Channel (SPACH). This is a forward link control channel that carries information for a specific mobile terminal.
  - (d) Random Access Channel (RACH). This is a reverse link channel that allows the mobile terminal to request access to the system.
2. Dedicated Traffic Channels (DTC). These are used to convey user information, as well as related control information between the base station and the mobile terminal. These channels are supported on both the forward link and the reverse link. The following traffic channels are specified in IS-136:
  - (a) Voice Traffic Channel. This channel carries compressed voice data between the base station and the mobile terminal.
  - (b) Fast Associated Control Channel (FACCH).
  - (c) Slow Associated Control Channel (SACCH).

For digital traffic channels, the transmitted data is divided into frames. Each frame has a duration of 40 msec and contains 1944 bits, such that the data rate is 48.6 kbps. The frame is divided into six time slots, each time slot is 6.67 msec and has 324 bits. A full rate channel occupies two time slots per frame as shown in Figure 1.9.a. While, a half rate channel occupies one time slot per frame as show in Figure 1.9.b.

Figure 1.10 shows the slot format for a forward digital traffic channel transmitted from the base station to the mobile terminal. The SYNC field is a 28 bit field used for synchronization and training. SACCH is the slow associated control channel. CDVCC is a Coded Digital Verification Color Code. RSVD is a reserved bit that is set to logical one. CDL is the Coded Digital Control Channel Locator. The voice data is separated into two fields each 130 bits wide.

Figure 1.11 shows the slot format for a reverse digital traffic channel transmitted from the mobile terminal to the base station. In addition to the



*Figure 1.9.* Frame structure for IS-136. (a) Full rate slot numbering. (b) Half rate slot numbering.

No. of Bits	SYNC	SACCH	Voice Data	CDVCC	Voice Data	Reserved	CDL
	28	12	130	12	130	1	11

*Figure 1.10.* Slot format for a forward link digital traffic channel.

SYNC field, the SACCH field and the CDVCC field which are similar to the forward channel, there are two extra-fields the guard time field and the ramp time field. Each mobile terminal is required to transmit data during the time slot assigned to it. Since the distance of the mobile terminal from the base station varies, hence the time of arrival of the time slots from different mobile terminals at the base station can vary. The guard time between time slots is to prevent consecutive slots arriving from different mobile terminals from overlapping. During the ramp time, the power of the mobile terminal increases from zero to its transmission power level. The voice data in the reverse channel time slot is split across three fields. The first field is 16 bits, while the second and third fields are 122 bits wide.

No. of Bits	Guard time	Ramp time	Voice Data	SYNC	Voice Data	SACCH	CDVCC	Voice Data
	6	6	16	28	122	12	12	122

Figure 1.11. Slot format for a reverse link digital traffic channel.

The voice data occupies 260 bits per time slot for both forward link and reverse link channels. The analog voice signal is sampled at 8 KHz. The Vector Sum Excited Linear Predictive coder compresses the sampled signal to 7.95 kbps. This corresponds to 159 bits every 20 msec, for a full rate voice channel. For half rate voice channels the compressed signal bit-rate is half this value. The 159 bits are split into 77 class 1 bits and 82 class 2 bits. The 12 most significant class 1 bits are protected by a 7-bit CRC. The 7-bit CRC along with the 77 class 1 bits are convolution encoded using a rate 1/2 convolution encoder with a constraint length of 5. The 178 coded class 1 bits are then concatenated with the 82 class 2 bits to generate the 260 bits of voice data. Figure 1.12 shows the sequence of voice coding operations for IS-136.

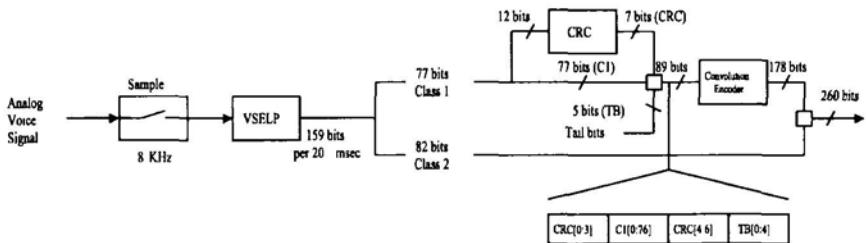


Figure 1.12. Coding of a voice channel in IS-136.

### 1.3.7 IS-95 North America CDMA Standard

The IS-95 standard is a dual mode cellular standard that supports both analog and digital communication modes. The analog mode of the IS-95 standard is based on the AMPS standard. While, the digital mode of the standard employs code division multiple access (CDMA). CDMA uses direct sequence spread spectrum as the multiple access technique for sharing the air interface among the different users. The CDMA cellular system uses variable rate speech coding based on the Code Excited Linear Predictive Coding (CELP)

algorithm. Power control is used in CDMA to guarantee that the received signal power of all users is approximately the same regardless of the location of the mobile terminal relative to the base station.

The frequency assignments for the forward link and the reverse link in IS-95 are identical to those of the AMPS standard. The forward link is allocated 25 MHz in the frequency band 869 MHz to 894 MHz. While, the reverse link is allocated 25 MHz in the frequency band 824 MHz to 849 MHz. Each CDMA channel has a chip rate of 1.2288 MHz and occupies a bandwidth of 1.25 MHz. This bandwidth is shared by more than one channel, channelization is achieved by using the orthogonal Walsh-Hadamard sequence as well as pseudonoise (PN) sequences.

In IS-95 there are four types of channels on the forward link (from the base station to the mobile terminal):

1. **Pilot Channel.** This channel is used to aid in the acquisition of the in-phase and quadrature-phase components on the PN sequence, and for coherent detection. The pilot channel is transmitted at a higher power than the other channels. The data transmitted on the pilot channel is all zero symbols. The pilot channel is assigned to Walsh sequence 0. The pilot channel accounts for 20% of the base station transmit power.
2. **Sync Channel.** This channel enables the mobile terminal to acquire frame synchronization with the base station. The sync channel is a 1.2 Kbps channel, with the data transmitted across it made up of frames, each having a period of 26.667 msec. Three frames make a super frame of period 80 msec. The data transmitted on the sync channel passes through a rate 1/2 convolution encoder, repeated twice and then interleaved and spread with a Walsh 32 function.
3. **Paging Channel.** This channel is used to page the mobile terminal and alert it when there is an incoming call at the base station. The paging channel is also used in response to an access channel message on the reverse link. The paging channel supports two data rates 9.6 Kbps and 4.8 Kbps. The data transmitted on the paging channel passes through a rate 1/2 convolution encoder and is then interleaved and spread with a Walsh function having an index of value 1 through 7.
4. **Traffic Channels.** These channels are used to convey voice data to the individual mobile terminals. The traffic channels support four data rates for the voice encoded signals: 8.6 Kbps, 4.0 Kbps, 2.0 Kbps, and 0.8 Kbps. The data transmitted on the forward link traffic channel is protected by CRC bits and the passes through a rate 1/2 convolution encoder. The output of the convolution encoder is repeated to a bit rate of 19.6 kbps, and is then interleaved and spread to the chip rate.

On the reverse link (from the mobile terminal to the base station) there are two types of channels:

1. **Access Channel.** This channel is used when the mobile terminal attempts to initiate a call, or in response to a paging message on the forward link. The data rate on the access channel before encoding is 4.0 Kbps. The data is encoded similar to the data encoding on the traffic channel of the reverse link.
2. **Traffic Channels.** These channels are used to convey voice data from the individual mobile terminals. The data rates supported on the reverse link as well as the data encoding is similar to that of the traffic channel of the forward link, except that it uses a rate 1/3 convolution encoder.

The reverse link PN sequence is generated using a polynomial of degree 42. The resulting PN-sequence should have good cross correlation properties, this allows channelization to be achieved by using different offsets (phases) of the PN sequence. In the forward link, the short PN code is used, this is generated from a polynomial of degree 15.

The advantages of CDMA include the ability to combat multi-path fading by using the RAKE receiver, as well as the ability to transmit variable data rates across the channel. However, CDMA does have its problems such as the near-far effect and unsynchronized codes on the reverse (up) link.

Figure 1.13 shows the block diagram of the forward link transmitter of a CDMA system. The output of the vocoder has a variable date rate that depends on voice activity. Table 1.3, shows the allowed data rates and number of bits per 20 msec frame, as well as the number of CRC bits for each data rate. A repeater is used after the convolution encoder, the repetition factor of the repeater is set such that its output has a constant bit rate. Before multiplying by the Walsh function, the signal is up sampled 64 times to a chip rate of 1.2288 Mcps.

*Table 1.3. Vocoder output rates.*

Vocoder output rate	Bits per frame	CRC bits
8.6 Kbps	172	12
4.0 Kbps	80	8
2.0 Kbps	40	0
0.8 Kbps	16	0

Figure 1.14 shows the block diagram of the reverse link transmitter of a CDMA system. The output of the vocoder has a variable date rate that depends on the voice activity as shown in Table 1.3. A rate 1/3 convolution encoder is used in the reverse link transmitter. This is followed by a repeater, the

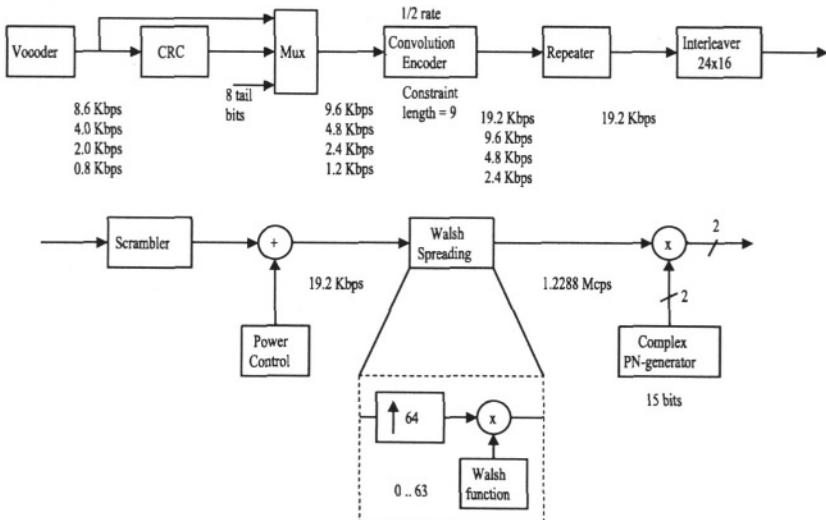


Figure 1.13. Signal processing in the forward link transmitter of a CDMA system.

repetition factor of the repeater is set such that its output bit rate is constant independent of the voice activity. The data is then interleaved, the output of the interleaver is grouped into 6-bit words, each word is mapped into a 64 bit Walsh sequence, hence the output of the Walsh modulator has a chip rate of 307.2 Kcps. This signal is up sampled 4 times to 1.2288 Mcps and multiplied by the output of a 42-stage complex PN generator.

## 1.4 TERMINOLOGY

**Mobile terminal.** This is the equipment that enables the user to communicate with another party. The information is conveyed to and from the mobile terminal through a wireless link. The mobile terminal is sometimes called the user equipment (UE).

**Base station.** This is the system that links the wireless air interface to the wired backbone network. The base station has a fixed location.

**Forward channel.** This is the channel between the base station and the mobile terminal. It is also known as the down-link.

**Reverse channel.** This is the channel between the mobile terminal and the base station. It is also known as the up-link.

**Simplex System.** This is a communication system that allows transmission in one direction only.

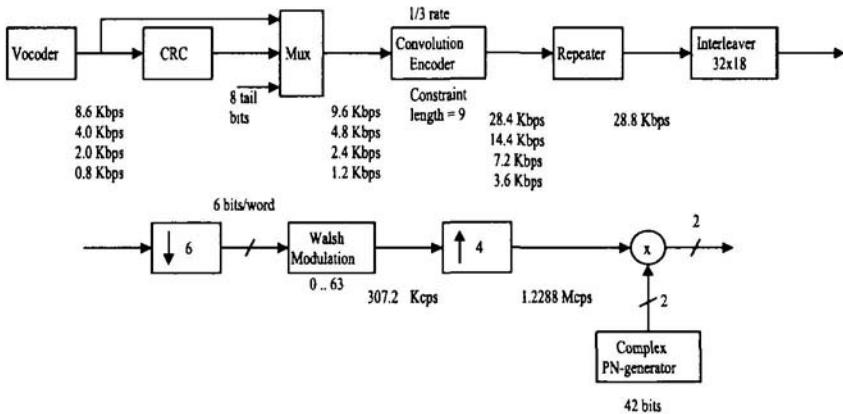


Figure 1.14. Signal processing in the reverse link transmitter of a CDMA system.

**Half Duplex System.** This is a communication system that allows transmission in both directions between two terminals. However, at a given time a terminal can only receive or transmit information. It can't do both.

**Full Duplex System.** This is a communication system that allows simultaneous transmission in both directions between two terminals. A terminal can transmit and receive information at the same time.

**Frequency division duplexing (FDD).** It is a duplexing technique that allows simultaneous radio transmission between the mobile terminal and base station by assigning two different frequencies for each communication session between the base station and the mobile terminal. One frequency is used for the forward radio channel, while the other frequency is used for the reverse radio channel.

**Time division duplexing (TDD).** It is a duplexing technique that allows the mobile terminal and base station to use a single frequency radio channel for transmission and reception, such that a portion of the time is used for the transmission of information from the base station to the mobile terminal and another portion of the time is used for the transmission of information from the mobile terminal to the base station. TDD is used for indoor and local area wireless applications with small coverage distances such as cordless phones.

**Handoff (inter-cell).** This is the process of transferring the mobile terminal from one base station to the next, as the user roams from one cell to another.

**Handoff (intra-cell).** This is the process of transferring the mobile terminal from one sector to another sector within the same cell.

**Hard Handoff.** This handoff process occurs abruptly, where the old wireless link is torn down and the new wireless link is brought up simultaneously.

**Soft Handoff.** As the mobile terminal moves from one cell to the other, a new wireless link to the new base station is brought up, co-existing with the old wireless link. Eventually, the old link is torn down.

**Mobile Switching Center (MSC).** It connects the base stations of its service area to the public switched telephone network (PSTN). The MSC controls the routing to calls to and from a mobile terminal, it also controls the handoff process of the mobile terminal from one base station to another.

**Adjacent Channel Interference (ACI).** This refers to the energy that spills over from one frequency band into an adjacent one.

**Co-channel Interference.** The same frequency band can be reused in a cellular system. The interference between two channels using the same frequency band is known as co-channel interference.

**Cell.** This is the coverage area of a base station.

**Received Signal Strength Indicator (RSSI).** This is the value of the received signal power, at the base station or the mobile terminal.

## 1.5 BOOK ORGANIZATION

The design of a cellular radio system, involves several engineering disciplines ranging from communication theory and digital signal processing to high frequency semiconductor technology and circuit design. In this book we cover these topics.

In chapter 2, we present an overview of wireless communication systems. We describe the block diagram of such a system and the function of each block. We also present two important limitations imposed on any wireless communication system, Nyquist criteria for inter-symbol interference free transmission, and the Shannon limit for the maximum information capacity in a noisy environment. We also consider the optimum receiver design.

In chapter 3, we discuss the characteristics of cellular systems, and the channel impairments that influence the design of such systems. In this chapter, we introduce the concepts of frequency reuse and handoff, we also discuss large scale and small scale fading models as well as the Doppler Shift. The chapter ends with a discussion of diversity techniques used to enhance the performance of wireless systems.

In chapter 4, we present the digital modulation techniques used in wireless systems. Modulation is the process of imparting the source information onto a sinusoidal carrier. Modulation schemes can be classified into amplitude modulation, frequency modulation and phase modulation, depending on which carrier parameter the information signal modulates. In chapter 4, we present

different digital modulation schemes, along with their power spectral density and bit error rate performance.

Spread spectrum is a wireless communication technique that transmits a signal in a bandwidth much larger than the bandwidth it occupies at baseband. This enables the transmitted signal to effectively overcome jamming, narrow-band interference and multi-path fading. Furthermore, multiple users can share the same bandwidth. The basic principles of spread spectrum are presented in chapter 5. In that chapter, we present the different types of spread spectrum techniques. We discuss the generation of pseudorandom noise sequences, which is essential in any spread spectrum system. Finally, we discuss some implementation details for spread spectrum systems such as synchronization, tracking and power control.

The performance of the receiver significantly impacts the performance of any wireless system. In chapter 6, we consider some of the receiver parameters that influence its performance, such as the noise figure and the intermodulation distortion. We also present several receiver architectures that are commonly used, such as the superheterodyne receiver and the homodyne receiver. Finally, we talk about software radio.

Analog-to-digital conversion is an essential operation in any digital transceiver to enable it to perform the signal processing operations in the digital domain. In chapter 7, we consider the parameters that characterize the performance of the analog-to-digital converter. We also discuss the operations that are needed to perform the analog-to-digital conversion, which include sampling and quantization. Next, we present different types of analog-to-digital converters, such as the flash, folding, interpolative, successive approximation, pipelining and Sigma-Delta analog-to-digital converters.

In chapter 8, we consider some practical issues pertaining to the design of a wireless transceiver. We begin by presenting the design constraints imposed on the transceiver. A wireless transceiver can be divided into two subsystems: the baseband subsystem, and the RF subsystem. In chapter 8, we consider the competing technologies for the implementation of each subsystem and the tradeoffs involved in selecting a certain technology.

The increase in demand for more functionalities and features in mobile terminals coupled with the limited energy batteries are capable of supplying, is making low-power design for wireless terminals an important design issue. In chapter 9, we consider the sources of power dissipation in digital circuits and present some design techniques developed to lower the power dissipation at the device and circuit levels, and at the architecture and algorithm levels.

Amplification is a fundamental signal processing operation performed in any wireless communication system. In chapter 10, we discuss the amplifier design issues for wireless systems. Amplification is need both in the receivers and transmitters. In the receiver, low noise amplifiers and automatic gain

control amplifiers are used. While in the transmitter power amplifiers are used. These amplifiers are presented in chapter 10.

Phase locked loops (PLLs) are an essential building block in wireless transceivers, where they are used for synchronization and carrier recovery. In chapter 11, we consider the structure and operation of PLLs. A PLL consists of several components such as: a phase detector, a frequency divider and a voltage controlled oscillator. In chapter 11, we consider the implementation and operation of each of these components. We also consider the theory of operation of oscillators along with some examples of oscillator circuits.

A frequency synthesizer is defined as an electronic device, which is capable of producing one or many frequencies from one or more reference sources. In chapter 12, we consider the design of frequency synthesizers, and present several important frequency synthesizer parameters. We also consider two different types of frequency synthesis techniques: the fractional-N frequency synthesizer and the direct digital frequency synthesizer. In that chapter, we also discuss the effect of phase noise on frequency synthesizers.

## Chapter 2

# WIRELESS COMMUNICATION SYSTEMS- OVERVIEW

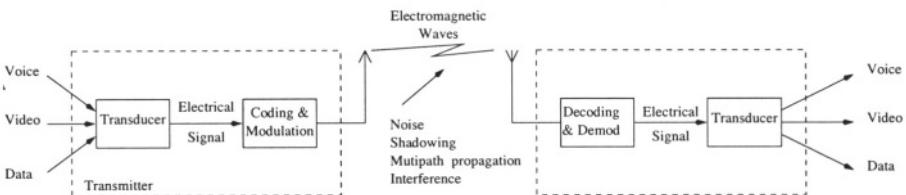
## 2.1 INTRODUCTION

This chapter deals with the basics of wireless communication systems. A wireless communications system conveys information from a source, which is known as the transmitter to a destination, which is known as the receiver. In section 2.2, we describe the basic elements of a wireless communications system and their functions. In section 2.3, we present the Nyquist criteria required for inter-symbol interference free transmission. In this section, we also describe a popular class of filters used in wireless systems known as raised cosine filters. In section 2.4, we describe the Shannon limit, which gives the information capacity of a noisy channel, we present some channel coding techniques used to achieve the Shannon limit. Finally, in section 2.5, we consider the optimum receiver design to minimize the bit error rate across an additive white Gaussian noise channel.

## 2.2 DIGITAL COMMUNICATION SYSTEMS

A wireless telecommunication system conveys information, such as voice, video or data from one location to another, by converting it into an electrical signal and eventually into an electromagnetic wave. Figure 2.1 shows a block diagram of such a system. At the transmitter side, a transducer converts the information signal (which could be a sound wave for example) into an electrical signal. The signal is then coded and modulated by the transmitter, eventually it is converted into an electromagnetic wave, and transmitted over a wireless channel to the receiver.

As the signal moves across the channel it suffers different types of impairments:



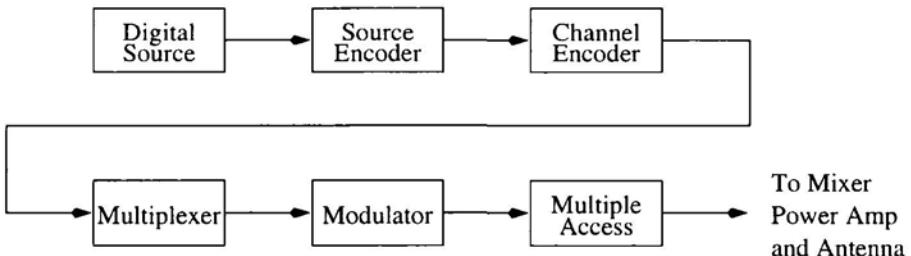
*Figure 2.1.* A wireless telecommunications system.

1. Attenuation.
2. Multi-path propagation.
3. Shadowing.
4. Doppler frequency shift.
5. Noise.
6. Interference (co-channel and adjacent channel).
7. Nonlinear distortion.

These channel impairments will be looked at in more detail in chapter 3. Eventually, the electromagnetic wave arrives at the receiver's antenna, which converts it back into an electrical signal. The electrical signal is demodulated and decoded, to identify it from the noise and interference it is embedded in. A transducer eventually transforms the electrical signal into its original form (sound wave for example).

The block diagram of a digital transmitter is shown in Figure 2.2. The source generates an electrical signal representing the information to be conveyed across the channel. The information carried by this electrical signal contains some irrelevancy and redundancy. Irrelevant information is unnecessary information beyond the perception of the human being. Redundant information is extra information, which can be removed without affecting the information content of the signal. The source encoder removes both irrelevant and redundant information.

The channel encoder adds some redundancy to the signal for error correction and error detection. Examples of error correction and error detection codes used in the channel encoder include block and convolution coding. Interleaving is also used in the channel encoder to enhance the performance of the system against bursty errors. If more than one source share a single channel, the data from the different sources needs to be multiplexed on to a single channel. The multiplexing can be done in the frequency domain or in the time domain.



*Figure 2.2.* Generic block diagram of a digital transmitter.

The signal is then modulated. This involves mapping each bit or symbol onto a certain waveform. Modulation techniques can be divided into three categories: amplitude modulation where the digital signal modulates the amplitude of the carrier, frequency modulation where the digital signal modulates the frequency of the carrier, and phase modulation where the digital signal modulates the phase of the carrier. Hybrid modulation techniques also exist. A distinguishing feature of frequency and phase modulations, is that ideally they both have a constant envelope. This has a two-fold advantage. First, a constant envelope reduces the non-linearities introduced by the power amplifier. Second, since no information is encoded in the amplitude, any amplitude variations due to noise or interference have a minimal effect on the signal-to-noise ratio performance of the system. Digital modulation techniques are presented in chapter 4.

The wireless channel has to be shared by many users. Various multiple access techniques exist to allow this sharing process to happen. For packet switched channels, techniques such as Carrier Sense Multiple Access (CSMA) and ALOHA are used for sharing the wireless channel. For circuit switched channels, where a physical channel is established during the entire communication session, techniques such as Frequency Division Multiple Access (FDMA), Time Division Multiple Access (TDMA), and Code Division Multiple Access (CDMA) are used for sharing the wireless channel.

Eventually, the transmitted signal is up converted by the mixer, amplified by the power amplifier and transmitted as an electromagnetic wave across the wireless channel.

The operations occurring at the receiver are the opposite of those occurring at the transmitter. Figure 2.3 shows the block diagram of a digital receiver. At the receiver side, the antenna converts the electromagnetic wave into an electrical signal. The electrical signal is amplified by the low noise amplifier (LNA), and down converted by the mixer.

The carrier frequency and the symbol timing are extracted for synchronization between the receiver and the transmitter. The multiple access block of the receiver determines the packets intended for this receiver in a packet switched

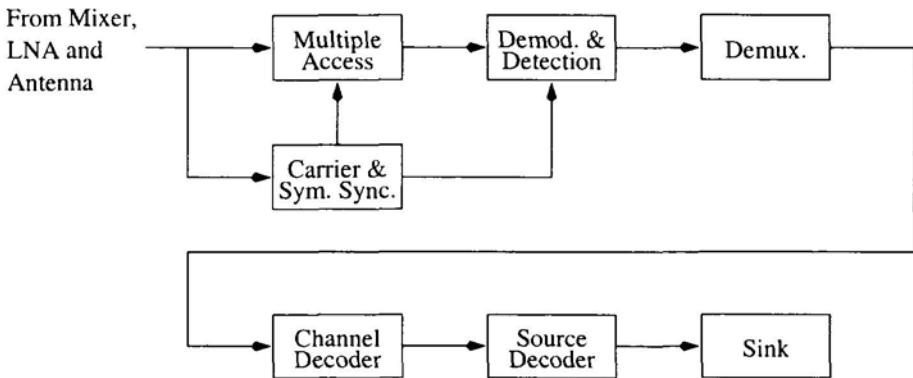


Figure 2.3. Generic block diagram of a digital receiver.

system, or the channel to be received in a circuit switched system. This channel can be a frequency band in an FDMA system, a time slot in a frame in a TDMA system, or a pseudonoise (PN) code in a CDMA system.

The demodulator block demodulates the received signal and detects the received symbol. Transmission over a wireless channels leads to the inevitable introduction of errors. These errors are corrected by the channel decoder, which also removes the redundancy that has been added by the channel encoder of the transmitter.

The source decoder undoes what the source encoder of the transmitter did, it adds any redundant information that has been removed by the source encoder. Irrelevant information that has been removed by the source encoder can not be reproduced by the source decoder. Hence, there will be some distortion between the information signal (sound or video) at that transmitter and that at the receiver. After the source decoder, the electrical signal can be transformed by a transducer into its original form.

The design of a wireless digital communications system involves many tradeoffs and compromises. The designer has a set of goals to achieve, with some limitations. In general, the goals of the designer include:

1. Maximizing the transmission bit rate,  $R$ , in a given bandwidth,  $W$ . Alternatively, this can be stated as minimizing the bandwidth,  $W$ , for a given transmission bit rate,  $R$ . The ratio  $R/W$  is known as the spectral efficiency. Therefore, one goal the designer seeks to achieve is maximizing the spectral efficiency of the system. This in turn leads to maximizing the utilization (number of users) of a wireless communications system.
2. Minimizing the bit error rate,  $P_e$ , for a certain signal energy to noise power spectral density ratio ( $E_b/N_o$ ). Where,  $E_b$  is the energy per bit,  $N_o$  is the noise power spectral density, both measured at the receiver. Alternatively,

the designer tries to minimize  $E_b/N_o$  for a certain bit error rate,  $P_e$ . This leads to higher power efficiency.

### 3. Minimize system complexity, and hence cost.

The limitations imposed on the designer of a wireless communication system are diverse. They range from physical limitations imposed by nature to man-made limitations such as government regulations and standards specifications, to technological limitations determined by the current state-of-the-art.

Physical limitations are absolute limits determined by nature. These include limits such as the Nyquist transmission rate which determines the maximum inter-symbol interference free transmission rate for a given channel bandwidth, this is presented in section 2.3. Another physical limitation is the Shannon limit which determines the maximum channel capacity for a given signal-to-noise ratio. The Shannon limit is presented in section 2.4.

Government regulations specify the frequency allocations of each system, the permissible emission power levels, minimum performance requirements of the receiver, etc. The goal of government regulations, is to allow different wireless systems to coexist without having excessive interference that could degrade the performance of any system. Government regulations and standards specifications also allow compatibility between the equipment of various manufacturers.

To reach the physical limits imposed by nature, intricate signal processing is required. The implementation of these complex operations into physically realizable hardware is limited by the state-of-the-art in technology and hardware design. As the level of integration and speed of operation of VLSI systems increase, more complex algorithms and operations, which were once unrealizable, are becoming more feasible to realize.

## 2.3 MINIMUM BANDWIDTH REQUIREMENT (THE NYQUIST LIMIT)

Consider a sampled signal given by:

$$x_s(t) = \sum_{n=-\infty}^{\infty} x_n \delta(t - nT_s) \quad (2.1)$$

Where,  $x_n$  represents the nth symbol. This sampled signal is passed through a filter having a transfer function  $H(f)$  and an impulse response  $h(t)$ . The output of the filter  $y_s(t)$  is related to the input of the filter and its impulse response by the following equation:

$$y_s(t) = x_s(t) * h(t)$$

$$= \sum_{n=-\infty}^{\infty} x_n h(t - nT_s) \quad (2.2)$$

Where,  $*$  denotes convolution. The output of the filter at  $t = 0$  is given by:

$$\begin{aligned} y_s(0) &= \sum_{n=-\infty}^{\infty} x_n h(-nT_s) \\ &= x_0 h(0) + \underbrace{\sum_{n \neq 0} h(-nT_s)}_{ISI} \end{aligned} \quad (2.3)$$

If the filter has zero delay, than  $x_0 h(0)$  is the desired output of the filter, this is the first term on the right-hand side of equation 2.3. The second term on the right-hand side of equation 2.3 is the interference of the adjacent symbols on the desired symbol, this term is called **Inter-symbol Interference (ISI)**.

For no inter-symbol interference to occur, the impulse response of the filter has to satisfy the following criteria:

$$h(nT_s) = \begin{cases} 1 & n = 0 \\ 0 & n \neq 0 \end{cases} \quad (2.4)$$

### 2.3.1 Nyquist Minimum Transmission Bandwidth Theorem

This theorem is also known as Nyquist's first theorem. It states that the minimum bandwidth of a low-pass filter that satisfies the no inter-symbol interference criteria of equation 2.4, for a sampled signal having a sample period  $T_s$  and a sampling frequency  $f_s = 1/T_s$  is  $f_N = f_s/2$  [8]. The filter that meets this requirement has the following transfer function:

$$H(f) = \begin{cases} T_s & |f| \leq f_N \\ 0 & |f| > f_N \end{cases} \quad (2.5)$$

The frequency  $f_N$  is known as the Nyquist frequency. This filter is sometimes referred to as the brick wall filter, its impulse response is given by:

$$h(t) = \frac{\sin(\pi t/T_s)}{(\pi t/T_s)} \quad (2.6)$$

Notice that, the impulse response of this filter does indeed satisfy the no Inter-symbol interference requirement of equation 2.4. However, because of the infinitely sharp slope in the transition from the pass-band to the stop-band

such a filter will require an infinite number of filtering sections to be realized. More importantly, because the impulse response of the brick wall filter decays with  $1/t$ , a particular choice of symbols together with a small sampling jitter can lead to very large inter-symbol interference [9].

There is a class of functions that satisfies the no inter-symbol interference criteria of equation 2.4 and yet don't have infinite sharpness. Let  $H(f)$  be one of these functions, with  $H(f)$  band-limited to  $f_s$ , i.e.

$$H(f) = 0 \quad \text{if } |f| > f_s$$

Assume that  $H(f)$  and its impulse response  $h(t)$  are real functions. Then both of them must also be even functions (have even symmetry around  $f = 0$  and  $t = 0$ ). For no inter-symbol interference,  $H(f)$  must satisfy the following condition:

$$H(f) + H(f_s - f) = T_s \quad 0 \leq |f| \leq \frac{f_s}{2} \quad (2.7)$$

Figure 2.4 shows three filters that satisfy the no inter-symbol interference criteria. Figure 2.4.a is the ideal brick wall filter. Figures 2.4.b and 2.4.c are filters that satisfy the condition of equation 2.7. The filters of Figure 2.4.b and 2.4.c are said to have vestigial symmetry.

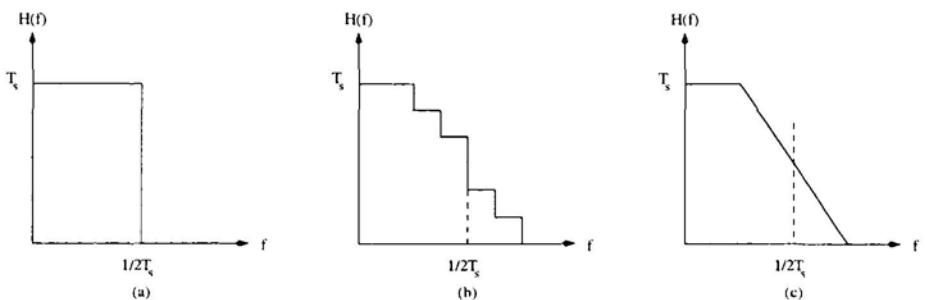


Figure 2.4. Inter-symbol interference-free filters. Filter (a) is the brick wall filter. Filters (b) and (c) satisfy condition 2.7.

### 2.3.2 The Raised Cosine Filter

A practical group of functions that satisfies condition 2.7 is the raised cosine filter which has a frequency response given by:

$$H(f) = \begin{cases} T_s & 0 \leq |f| \leq \frac{1-\alpha}{2T_s} \\ \frac{T_s}{2} \left\{ 1 + \cos \left[ \frac{\pi T_s}{\alpha} \left( |f| - \frac{1-\alpha}{2T_s} \right) \right] \right\} & \frac{1-\alpha}{2T_s} < |f| \leq \frac{1+\alpha}{2T_s} \\ 0 & |f| > \frac{1+\alpha}{2T_s} \end{cases} \quad (2.8)$$

Where,  $0 \leq \alpha \leq 1$ . When  $\alpha = 0$ , the raised cosine filter becomes the brick wall filter of equation 2.5. When  $\alpha = 1$ , the transfer function of the raised cosine filter becomes:

$$H(f) = \begin{cases} T_s \cos^2 \left[ \frac{\pi T_s f}{2} \right] & |f| \leq \frac{1}{T_s} \\ 0 & \text{otherwise} \end{cases} \quad (2.9)$$

The bandwidth of the raised cosine filter,  $W$ , is given by:

$$W = \frac{f_s}{2}(1 + \alpha) = f_N(1 + \alpha) \quad (2.10)$$

The parameter  $\alpha$ , known as the roll-off factor, determines the excess bandwidth relative to the minimum bandwidth of the brick wall filter. When  $\alpha = 1$ , the bandwidth of the raised cosine filter is double the minimum bandwidth of the brick wall filter. The impulse response of the raised cosine filter is given by:

$$h(t) = \frac{\sin(\pi t/T_s)}{(\pi t/T_s)} \left[ \frac{\cos(\alpha\pi t/T_s)}{1 - (2\alpha t/T_s)^2} \right] \quad (2.11)$$

Figure 2.5 shows the transfer function of the raised cosine filter for different roll-off factors. The impulse response of the raised cosine filter is shown in Figure 2.6. Notice that, the impulse response of the raised cosine filter decays faster as  $\alpha$  gets larger, which makes it more robust with respect to sampling jitters. The noise bandwidth  $\text{BW}_n$  of the raised cosine filter is given by:

$$\text{BW}_n = (1 - \frac{\alpha}{4}) \frac{f_s}{2} \quad (2.12)$$

It should be noted that the raised cosine filter, whose transfer function is given by equation 2.8 and whose impulse response is given by equation 2.11, is a non-causal filter, which means that the output of the filter exists before the arrival of the input signal, such a filter is not a realizable filter. However, the raised cosine filter can be implemented by truncating its impulse response to a finite time and then shifting it, by introducing a delay, such that the impulse

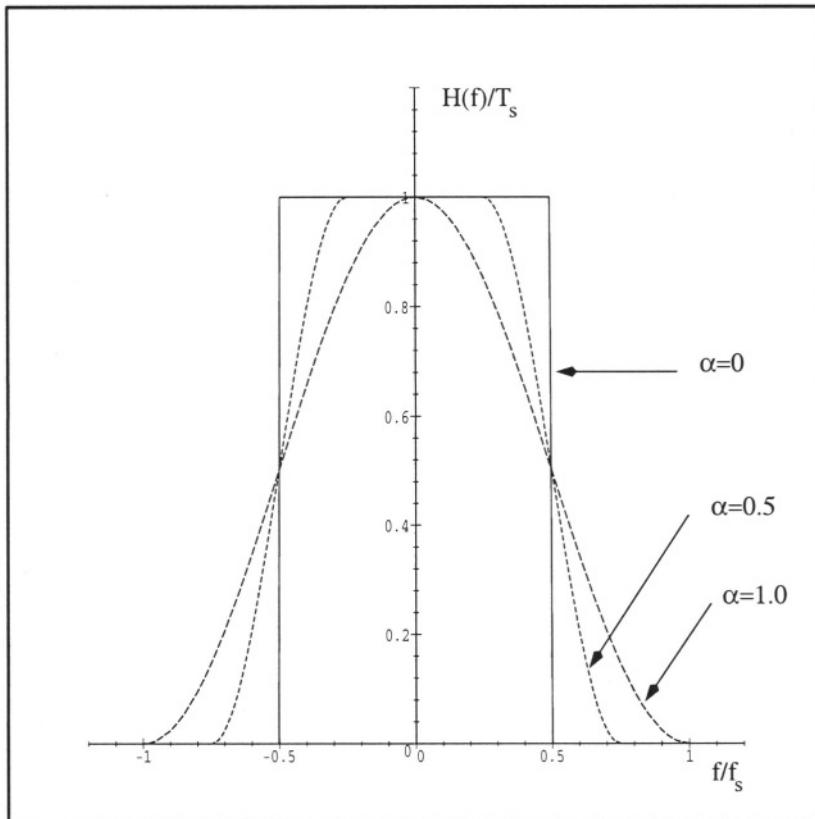


Figure 2.5. The transfer function of a raised cosine filter with different roll-off factors,  $\alpha = 0.0, 0.5$  and  $1.0$ .

response exists for positive time only. Figure 2.7 shows the impulse response of a realizable approximation to the raised cosine filter. The raised cosine response is truncated to  $6T_s$  and shifted (delayed) by  $3T_s$ .

In section 2.5, it will be demonstrated that the optimum filter, with respect to bit error rate performance for a given signal to noise ratio is the matched filter. This necessitates the use of two identical filters at the transmitter and receiver. The combined frequency response of the transmitter filter, receiver filter along with the frequency response of the channel should be a raised cosine filter that guarantees no inter-symbol interference. If we assume that the frequency response of the channel is flat over the band of interest, then the filters used in the transmitter and receiver should be the square root of the

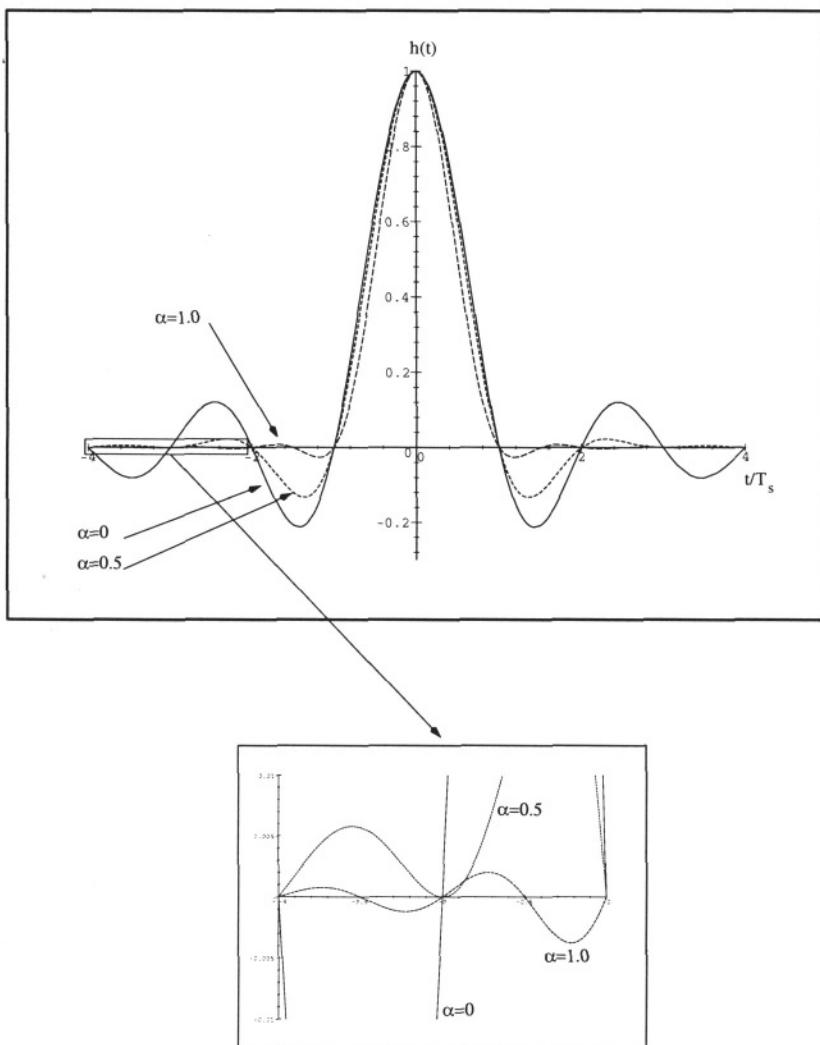


Figure 2.6. The impulse response of a raised cosine filter with different roll-off factors,  $\alpha = 1, 0.5$  and  $1$ .

raised cosine filter, hence called a square root raised cosine filter which has a frequency response given by:

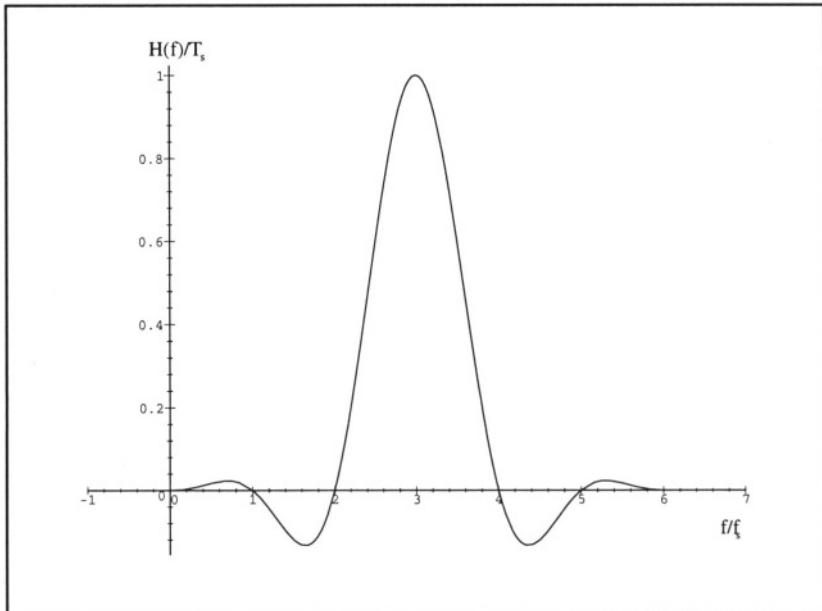


Figure 2.7. The impulse response of a realizable approximation of the impulse response of a raised cosine filter. The impulse response is truncated to  $6T_s$  and shifted (delayed) by  $3T_s$ . The raised cosine filter has a roll-off factor  $\alpha = 0.5$ .

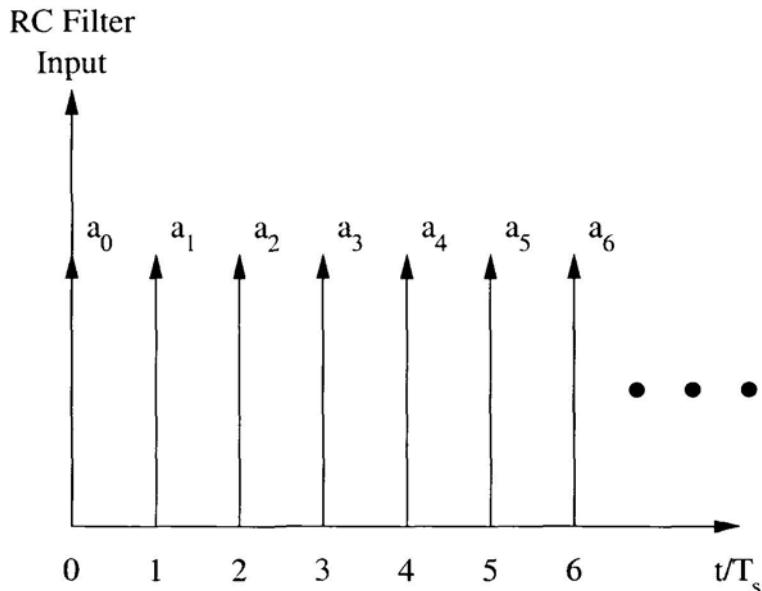
$$H(f) = \begin{cases} \sqrt{T_s} & 0 \leq |f| \leq \frac{1-\alpha}{2} \\ \sqrt{T_s} \cos \left[ \frac{\pi T_s}{2\alpha} \left( |f| - \frac{1-\alpha}{2T_s} \right) \right] & \frac{1-\alpha}{2T_s} < |f| \leq \frac{1+\alpha}{2T_s} \\ 0 & |f| > \frac{1+\alpha}{2T_s} \end{cases} \quad (2.13)$$

The impulse response of the square root raised cosine filter is given by:

$$h(t) = 4\alpha \frac{\cos \left[ \frac{(1+\alpha)\pi t}{T_s} \right] + \left( \frac{4\alpha t}{T_s} \right)^{-1} \cdot \sin \left[ \frac{(1-\alpha)\pi t}{T_s} \right]}{\pi \sqrt{T_s} \left[ \left( \frac{4\alpha t}{T_s} \right)^2 - 1 \right]} \quad (2.14)$$

The noise bandwidth of square root raised cosine filters is given by:

$$\text{BW}_n = \frac{f_s}{2} \quad (2.15)$$



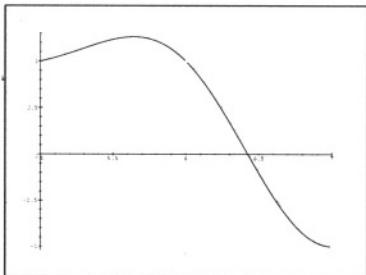
*Figure 2.8.* The input impulse sequence to a raised cosine filter having a truncated and delayed impulse response as shown in Figure 2.7.  $a_i$  can be 1 or -1.

**Example 2.1.** Assume that the sequence of impulses shown in Figure 2.8 are injected to the input of a raised cosine filter having a roll-off factor  $\alpha = 0.5$ , and a truncated and delayed impulse response as shown in Figure 2.7.  $a_i$  is 1 for a logic 0 and -1 for a logic 1. Draw the output waveform of the raised cosine filter between times  $5T_s$  and  $7T_s$ , for the following input sequences:

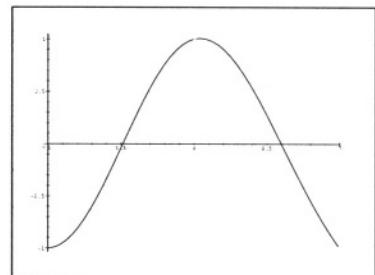
- (a) 0 0 0 0 1 0 0.
- (b) 1 0 1 0 1 1 0.
- (c) 0 1 1 0 1 0 0.
- (d) 1 1 0 0 1 0 1.
- (e) 1 1 101 1 1.

For each case find the zero crossing that exists between  $6T_s$  and  $7T_s$ .

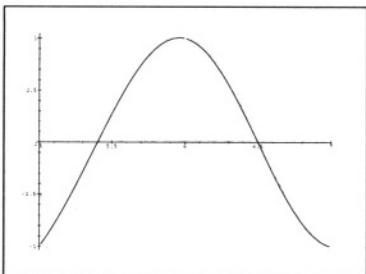
**Solution.** The RC filter causes a delay of  $3T_s$ , hence the output of the filter at a time  $5T_s$  corresponds to the second sample  $a_2$ , and the output of the filter at time  $7T_s$  corresponds to the fourth sample  $a_4$ . Figure 2.9 shows the output of the raised cosine filter for each input sequence. The time of the zero crossing between  $6T_s$  and  $7T_s$  is given in Table 2.1.



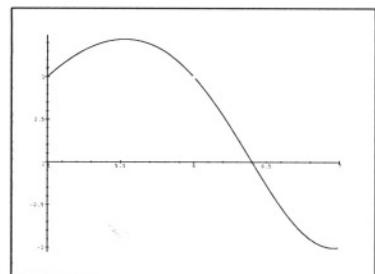
(a)



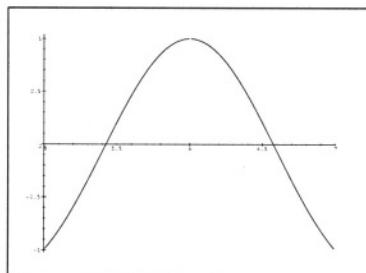
(b)



(c)



(d)



(e)

Figure 2.9. The output response of a raised cosine filter having a truncated and delayed impulse response for the following input sequences: (a) 0 0 0 0 1 0 0. (b) 1 0 1 0 1 1 0. (c) 0 1 1 0 1 0 0. (d) 1 1 0 0 1 0 1. (e) 1 1 1 0 1 1 1.

As can be seen from Figure 2.9 and Table 2.1, the position of the zero crossing between two samples with opposite polarity depends on the input sequence. The peak-to-peak deviation between the zero crossing is known as the data transition jitter,  $J_{PP}$ , [10]. If  $J_{PP}$  is too large, it can have a detrimental

Table 2.1. Zero crossing time between  $6T_s$  and  $7T_s$  for various input sequences to a raised cosine with a truncated and delayed impulse response as given in Figure 2.7.

Input Sequence	Zero Crossing Time
0 0 0 0 1 0 0	$6.425T_s$
1 0 1 0 1 1 0	$6.6T_s$
0 1 1 0 1 0 0	$6.5T_s$
1 1 0 0 1 0 1	$6.4T_s$
1 1 1 0 1 1 1	$6.575T_s$

effect on the performance of the symbol timing recovery circuit and hence, on the overall system performance.

The raised cosine filter with a roll-off factor  $\alpha = 1.0$  provides jitter-free performance. The impulse response of this filter is given by:

$$h(t) = \frac{\sin(\pi t/T_s)}{(\pi t/T_s)} \frac{\cos(\pi t/T_s)}{(1 - (2t/T_s)^2)} \quad (2.16)$$

Notice that, in addition to the zeros at  $\pm T_s$ ,  $\pm 2T_s$  . . . , there are new zeros at  $\pm 0.5T_s$ ,  $\pm 1.5T_s$ , . . . . By introducing zeros midway between the sampling points, the filter provides jitter-free performance. As the roll-off factor,  $\alpha$ , starts to decrease from 1 and approaches 0,  $J_{pp}$  starts to increase, as can be seen in Figure 2.10.

## 2.4 THE SHANNON LIMIT

In 1948, Claude Shannon developed the theoretical bound on error correction codes. This limit has been known as the Shannon limit. Shannon's limit gives the theoretical limit of information rate that can be carried by a channel error-free, using the most optimum error correction code. This limit is given by [11]:

$$C = BW \log_2 \left( 1 + \frac{S}{N} \right) \quad (2.17)$$

Where,

$C$  is the channel information capacity in bits per second.

$BW$  is the channel bandwidth in Hz.

$S/N$  is the signal-to-noise ratio (SNR).

The SNR is related to the bit energy and the noise power spectral density by the following equation:

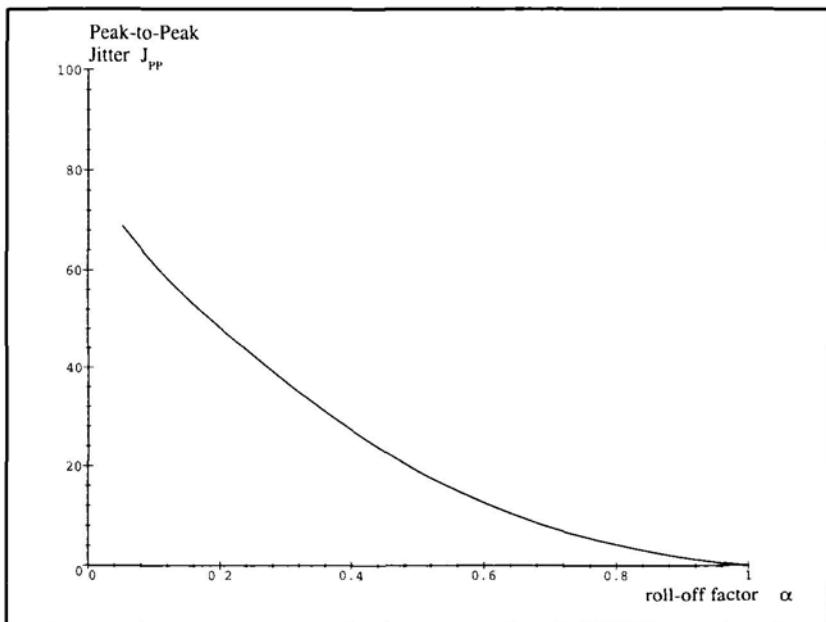


Figure 2.10. The peak-to-peak jitter,  $J_{pp}$ , versus the roll-off factor,  $\alpha$ , of a raised cosine filter.

$$\text{SNR} = \frac{E_b R_b}{N_o \text{BW}} \quad (2.18)$$

Where,

$E_b$  is the energy per bit of the received signal,

$R_b$  is the data transmission rate in bits per second,

$N_o$  is the noise power spectral density in Watts/Hz.

The channel capacity can be increased by increasing the channel bandwidth or by increasing the signal-to-noise ratio. Hence, Shannon's theorem allows us to compromise between the bandwidth efficiency of the channel, i.e. how many bits per second are carried per 1 Hz bandwidth, and the signal-to-noise ratio of the channel, to achieve a certain error-free transmission rate. Shannon's theorem doesn't tell us how to achieve this error-free transmission, it only tells us that it is possible. To achieve error-free (or almost error-free) transmission complex and elaborate coding techniques are used.

For the channel capacity to approach Shannon's limit forward error correction (FEC) channel coding is used. Channel coding introduces redundancy into

the transmitted bit stream, such that only a subset of all possible bit patterns is transmitted. This redundancy enables the receiver to estimate the errors due to noise, by approximating the received bit pattern to one of the allowable code-words, and hence recovering the original bit stream that was transmitted.

In this section we present the fundamental concepts of channel coding. For more details, the reader can refer to [12] and [13]. In this section, we present linear block codes, convolution codes, as well as interleaving.

### **2.4.1 Linear Block Codes**

Assume an uncoded binary data stream. This data stream is divided into blocks. Each block has  $k$  bits. An  $(n,k)$  block code maps each  $k$ -bit uncoded message bits into an  $n$ -bit code-word, where,  $n > k$ . Thus, an  $(n,k)$  block code has  $n - k$  redundant check bits. An  $(n,k)$  block code is said to have a code rate  $k/n$ .

At the receiver, the decoder examines the received code-word. Since, not all  $n$ -bit code-words are transmitted (because of the redundancy) the decoder approximates the received code-word to the nearest allowable  $n$ -bit code-word and is thus able to provide error free reception at the output of decoder in the presence of a limited number of transmission errors at its input. The greater the “difference” between the  $n$ -bit code-words, the more tolerant the decoder to transmission errors.

The difference between any two  $n$ -bit code-words is defined as the number of bits the code-words differ in. This difference is known as the Hamming distance  $d_H$ . For the following 7-bit code-words: 0001011 and 0011101, the Hamming distance between them is 3, i.e. the code words differ in three bit position. If we were to transmit the first code-word (0001011) and an error were to occur in the second bit from the right, i.e. the received word is 0001001, the receiver calculates the Hamming distance between the received word and the allowable code-words. The Hamming distance is one for the former and two for the latter. Thus, the receiver approximates the received code word to the first code word and eliminates the error. However, in this example if an error were to occur in the second and third bit positions from the right, the decoder will not be able to recover the correct code-word.

The previous discussion indicates that a good block code needs to be designed such that the minimum Hamming distance between any two allowable code-words is as large as possible. This in turn maximizes the number of errors the code can correct.

A systematic block code consists of  $n$ -bit code-words whose first or last  $k$  bits are identical to the uncoded message bits. The remaining  $n - k$  bits are calculated to maximize the minimum Hamming distance between any two code-words.

Consider a  $(7, 4)$  linear systematic block code. Associated with this code is a  $4 \times 7$  generation matrix that maps each 4-bit uncoded message into a 7-bit code-word. The generation matrix for this code is given by:

$$G = \begin{bmatrix} 1 & 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix} \quad (2.19)$$

The 7-bit code-word is calculated from the 4-bit uncoded message block using the following equation:

$$C = MG \quad (2.20)$$

Where,  $C$  is the 7-bit code-word and  $M$  is the 4-bit uncoded message block. Notice that, the last three bits are the check bits and these are calculated according to the following equations:

$$c_5 = m_1 \oplus m_2 \oplus m_3 \oplus 0$$

$$c_6 = 0 \oplus m_2 \oplus m_3 \oplus m_4$$

$$c_7 = m_1 \oplus m_2 \oplus 0 \oplus m_4$$

Table 2.2 shows the 4-bit message word and the associated 7-bit code-word. For each code-word, there exists 7 code-words with a Hamming distance of 3, and 7 code-words with a Hamming distance of 4 and one code-word with a Hamming distance of one. Hence, the  $(7, 4)$  block code can correct up to one transmission error per block.

To decode the received word, the Hamming distance between the  $n$ -bit received word and each allowable  $n$ -bit code-word is calculated. The code-word that gives the minimum Hamming distance is then selected. Instead of doing an exhaustive search to find the closest code-word it is possible to multiple the received  $n$ -bit vector by a matrix, that depends on the block code used. The result of this multiplication is a vector known as the Syndrome. This vector points to the bit(s) that are in error. For more details about this algorithm refer to [14].

## 2.4.2 Convolution Codes

Convolution codes are designed to continuously encode each  $k$ -bit message word into an  $n$ -bit code-word, without dividing the input bit stream into distinct  $k$ -bit blocks. Convolution codes possess some memory, which depends on the

Table 2.2. Code-words for a (7,4) Systematic Block Code.

Message Word	Code-word
0 0 0 0	0 0 0 0 0 0 0
0 0 0 1	0 0 0 1 0 1 1
0 0 1 0	0 0 1 0 1 1 0
0 0 1 1	0 0 1 1 0 1 1
0 1 0 0	0 1 0 0 1 1 1
0 1 0 1	0 1 0 1 1 0 0
0 1 1 0	0 1 1 0 0 0 1
0 1 1 1	0 1 1 1 0 1 0
1 0 0 0	1 0 0 0 1 0 1
1 0 0 1	1 0 0 1 1 1 0
1 0 1 0	1 0 1 0 0 1 1
1 0 1 1	1 0 1 1 0 0 0
1 1 0 0	1 1 0 0 0 1 0
1 1 0 1	1 1 0 1 0 0 1
1 1 1 0	1 1 1 0 1 0 0
1 1 1 1	1 1 1 1 1 1 1

constraint length of the code. The convolution code is defined by its rate  $k/n$ , its constraint length  $l$  and the polynomials used to generate the output.

To understand the operation of convolution codes, lets consider a rate 1/2 code, with a constraint length of 3 and having two code generation polynomials given by:

$$G_1(X) = X^2 + X + 1$$

$$G_2(X) = X^2 + 1$$

Figure 2.11 shows the block diagram, of a rate-1/2 convolution encoder, having a constraint length of 3. The constraint length  $l$  defines the delay between the earliest and latest bits that can affect the output of the encoder. The number of delay stages in the convolution encoder is equal to  $l - 1$ . Since the convolution encoder has delay elements to hold  $l - 1$  of the previous bits of the input bit stream, its operation can be described by a finite state machine with  $2^{l-1}$  states as shown in Figure 2.12. Based on the value of the input bit to the encoder, the state of the finite state machine changes as shown in Figure 2.12. Associated with each branch of the finite state machine is a label that defines the input bit and the two output bits for the rate-1/2 convolution encoder.

Alternatively, the convolution encoder can be described by a trellis diagram as shown in Figure 2.13. Each stage has four nodes which correspond to the four states of the finite state machine of Figure 2.12. Every time the encoder

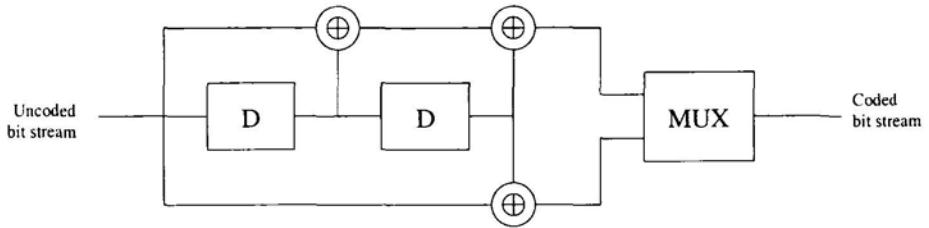


Figure 2.11. A rate-1/2 convolution encoder, with a constraint length of 3.

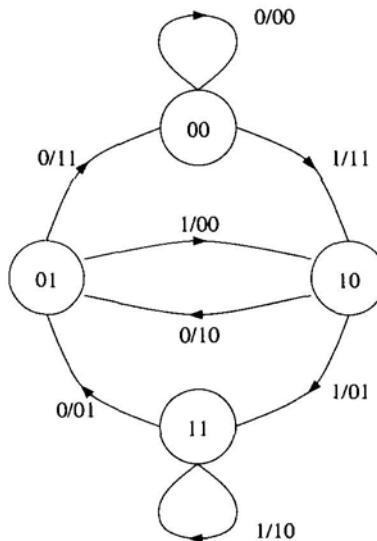


Figure 2.12. The finite state machine representation of the convolution encoder of Figure 2.11.

gets an input bit, it moves from one stage to the next, moving from left to right. The solid line represents a state transition when input is logic 0, while a dashed line represents a state transition when the input is logic 1.

There are several ways to decode a convolution-encoded bit stream, the most commonly used is the Viterbi algorithm. This is a maximum likelihood decoder. The idea is to use the trellis of Figure 2.13 and to find a path through the trellis that has the least Hamming distance to the received coded bit stream. Tracing back along this path generates the uncoded bit stream at the output of the decoder.

### 2.4.3 Interleaving

Interleaving doesn't add any redundancy to the bit stream, rather it changes the order the bits are transmitted across the channel, such that two consecutive bits

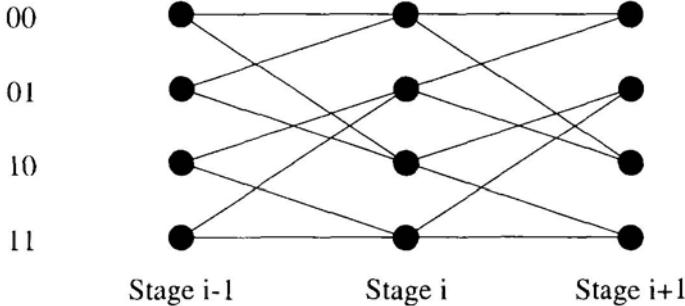


Figure 2.13. The trellis diagram of the convolution encoder of Figure 2.11.

are not transmitted consecutively across the channel. The reason for doing this is that when ever a fade occurs or a noise burst occurs, it usually covers multiple bit intervals. This leads to a burst of errors, where more than one consecutive bit is in error. Channel coding, whether block coding or convolution coding, is most optimum when the errors occur randomly and not in bursts. Interleaving can be viewed as a type of time diversity.

The interleaver is placed after the channel encoder in the transmitter and a de-interleaver is placed before the channel decoder in the receiver as shown in Figure 2.14. The de-interleaver reverses the interleave operation. In this case, should a burst of errors occur across the channel, the de-interleaver disperses these errors throughout the bit stream at the input to the channel decoder, making the errors appear as random errors to the channel decoder.



Figure 2.14. Interleaving and De-interleaving.

There are two types of interleavers that are commonly used, block interleavers and convolution interleavers. Figure 2.15 shows the block diagram of a block interleaver and de-interleaver. The block interleaver uses a RAM (Random Access Memory) having  $M$  rows and  $N$  columns. The input bit stream to the interleaver is written row by row starting with the top row. When all the  $M \times N$  memory blocks are written to, the data stored in the RAM is readout column by column, starting with the top element of the left-hand side column.

The de-interleaver reverses the interleaving operation. This is done by writing the received data at the input of the de-interleaver to a RAM having  $M$  rows and  $N$  column. The input bit stream to the de-interleaver is written

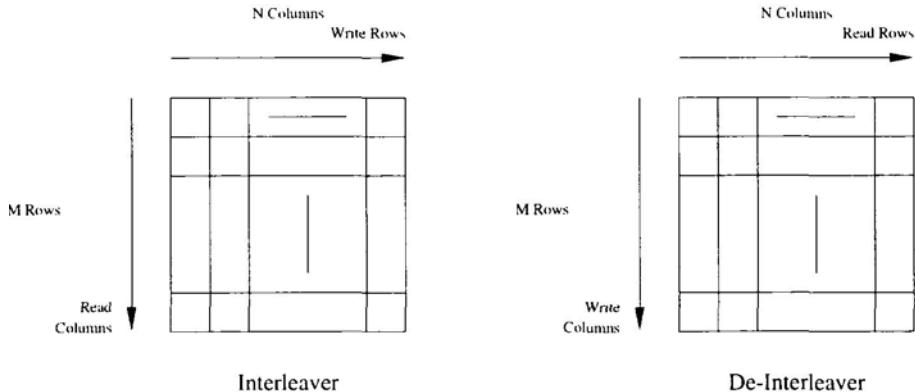


Figure 2.15. Block Interleaving and de-interleaving.

column by column starting with the left-hand side column. When all the  $M \times N$  memory blocks are written to, the data stored in the RAM is readout row by row starting with the top row.

**Example 2.2.** Consider a block interleaver having 4 rows and 3 columns. Assume that the input bit stream to the interleaver is:

$$x_1, x_2, \dots$$

Find the first 12 outputs of the interleaver.

**Solution.** The inputs to the interleaver are written row by row and hence they are arranged as follows:

$$\begin{bmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \\ x_{10} & x_{11} & x_{12} \end{bmatrix}$$

The outputs from the interleaver are read column by column, hence the output bit stream is:

$$x_1 x_4 x_7 x_{10} x_2 x_5 x_8 x_{11} x_3 x_6 x_9 x_{12}$$

A block interleaver operates on discrete blocks with no overlapping between blocks. Another type of interleaving is convolution interleaving. In this case, interleaving is more continuous as there is no clear boundary between one block and the next. Figure 2.16 shows a block diagram of a convolution interleaver. It consists of two synchronized switches and  $M$  parallel rows each

having a different delay. The first row has zero delay, the second has a delay  $N$ , the third has a delay  $2N$  and so on. i.e. successive interleaver inputs are delayed by different amounts before being transmitted across the channel

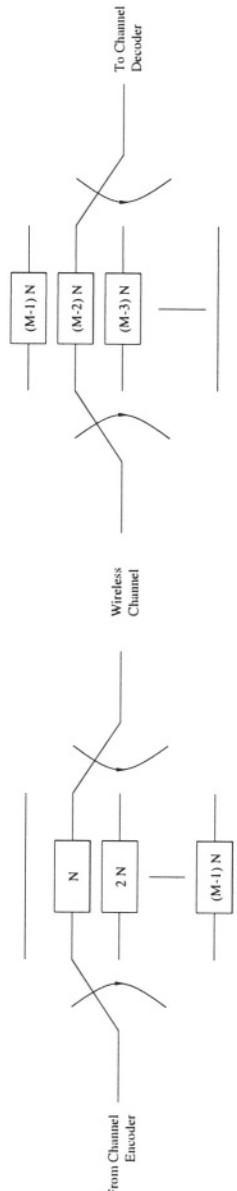


Figure 2.16. Convolution Interleaver and De-interleaver.

The de-interleaver reverses the interleaving operation. It too has two synchronized switches, that are also synchronized to those of the transmitter, and it has  $M$  parallel rows. The delays of each row are adjusted in such a way that the overall delay of any bit passing through the interleaver and the de-interleaver is the same.

## 2.5 OPTIMUM RECEIVER DESIGN FOR BAND-PASS DIGITAL SYSTEMS

In this section, we consider the design of a receiver that minimizes the probability of error in the presence of additive white Gaussian noise (AWGN). Assume that the wireless channel has infinite bandwidth, and consequently has no inter-symbol interference. Furthermore, assume that the digital signal  $b_k$  transmitted across the channel is a binary signal. The output of the modulator,  $s(t)$ , depends on the binary bit that is being modulated and generates one of two output waveforms according to the following equation:

$$s(t) = \begin{cases} s_0(t) & \text{if } b_k = 0 \\ s_1(t) & \text{if } b_k = 1 \end{cases} \quad (2.21)$$

One such waveform is transmitted every  $T_b$  seconds. Where,  $T_b$  is the bit duration. The signal  $s_i(t)$  is a finite energy signal having an energy:

$$E_{b_i} = \int_0^{T_b} s_i^2(t) dt \quad (2.22)$$

The transmitted signal passes through the channel whose model is shown in Figure 2.17. This channel has the following properties:

1. No gain or attenuation.
2. Infinite bandwidth.
3. Additive White Gaussian Noise (AWGN) with power spectral density (PSD)  $N_o$ .

Hence, the received signal can be expressed as:

$$r(t) = s(t) + n(t) \quad (2.23)$$

Where,  $n(t)$  is a sample function of a white Gaussian random process  $N(t)$ .

The received signal passes through a filter  $H_R(f)$  and is than sampled once every  $T_b$  seconds. The sampled output is compared against a threshold level to determine which waveform and consequently which bit was transmitted. The goal of the following analysis is to determine the optimum filter and the optimum threshold level needed to minimize the probability of bit error for a

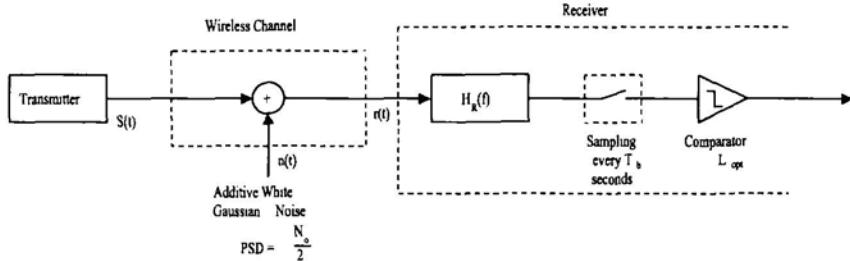


Figure 2.17. Additive white Gaussian noise channel model.

certain noise level. The output of the receiver filter  $H_R(f)$  can be expressed as:

$$x(t) = \int_{-\infty}^{\infty} r(\tau)h_R(t - \tau)d\tau \quad (2.24)$$

Where,  $h_R(t)$  is the impulse response of the receiver's filter. Assuming that  $h_R(t)$  is time limited, such that it has a non-zero value only during the interval  $[0, T_b]$ . hence the output of the receiver's filter at the sampling instance  $T_b$  can be expressed as:

$$\begin{aligned} x(T_b) &= \int_0^{T_b} r(\tau)h_R(T_b - \tau)d\tau \\ &= \int_0^{T_b} s(\tau)h_R(T_b - \tau) + \int_0^{T_b} n(\tau)h_R(T_b - \tau) \end{aligned} \quad (2.25)$$

The first term on the right-hand side of equation 2.25 is a deterministic value, which depends solely on whether  $s_1(t)$  or  $s_2(t)$  was transmitted.

$$V_i = \int_0^{T_b} s_i(\tau)h_R(T_b - \tau)d\tau \quad (2.26)$$

In the frequency domain:

$$V_i(f) = H_R(f)S_i(f) \quad (2.27)$$

Hence,

$$V_i = v_i(T_b) = \int_{-\infty}^{\infty} H_R(f)S_i(f)e^{j2\pi fT_b}df \quad (2.28)$$

The second term on the right-hand side of equation 2.25 is a Gaussian random variable that has a zero mean, and a variance  $\sigma_N^2$ .

$$\begin{aligned}\sigma_N^2 &= E \left[ \left( \int_0^{T_b} n(\tau) h_R(T_b - \tau) d\tau \right)^2 \right] \\ &= E \left[ \int_0^{T_b} \int_0^{T_b} n(u) n(v) h_R(T_b - u) h_R(T_b - v) du dv \right] \quad (2.29)\end{aligned}$$

Where,  $E[X]$  is the expected value of random variable  $X$ . Interchanging the order of integration and expectation we get:

$$\begin{aligned}\sigma_N^2 &= \int_0^{T_b} \int_0^{T_b} E[n(u)n(v)] h_R(T_b - u) h_R(T_b - v) du dv \\ &= \int_0^{T_b} \int_0^{T_b} R_N(u, v) h_R(T_b - u) h_R(T_b - v) du dv \\ &= \frac{N_o}{2} \int_0^{T_b} h_R^2(u) du \\ &= \frac{N_o}{2} \int_{-\infty}^{\infty} |H_R(f)|^2 df \quad (2.30)\end{aligned}$$

Where,  $R_N(u, v)$  is the autocorrelation function of the white Gaussian random process  $N(t)$ :

$$R_N(u, v) = \frac{N_o}{2} \delta(u - v) \quad (2.31)$$

The conditional probability distribution function of the received signal, given the transmitted waveform, is a Gaussian random variable with mean  $V_i$  and variance  $\sigma_N^2$ :

$$f(x/s_i) = \frac{1}{\sqrt{2\pi\sigma_N^2}} e^{-\frac{(x-V_i)^2}{2\sigma_N^2}} \quad (2.32)$$

$f(x/s_i)$  is known as the likelihood function. At the receiver a decision needs to be made on which waveform was transmitted based on the output of the receiver's filter  $x(T_b)$ . Assume that  $P(s_i/x)$  is the probability that waveform  $s_i(t)$  was transmitted given that the output of the receiver filter is  $x$ . The optimum decision rule selects the waveform that has the highest probability. This decision rule is known as maximum a posteriori (MAP) probability rule. Using Bayes' theorem:

$$P(s_i/x) = \frac{f(x/s_i)P(s_i)}{f(x)} \quad (2.33)$$

Where,  $P(s_i)$  is the probability that waveform  $s_i(t)$  was transmitted. Assuming equi-probable data streams, i.e.  $P(s_i) = 0.5$ .  $f(x)$  is the probability density function of the sampled output of the receiver's filter.

$$f(x) = \sum_i f(x/s_i)P(s_i) \quad (2.34)$$

Since  $f(x)$  is independent of  $s_i(t)$ , and assuming equi-probable data streams, maximizing  $P(s_i/x)$  is equivalent to maximizing  $f(x/s_i)$ . The optimum decision criteria is which input waveform, given that it was transmitted, maximizes the probability density function of the output waveform. This decision criteria is known as the maximum likelihood decision.

Figure 2.18 shows the conditional probability density function at the sampled output of the receiver given the transmitted waveform. Starting from equation 2.32 we can show that the optimum threshold level  $L_{opt}$  used in the decision block of Figure 2.17 is:

$$L_{opt} = \frac{V_1 + V_2}{2} \quad (2.35)$$

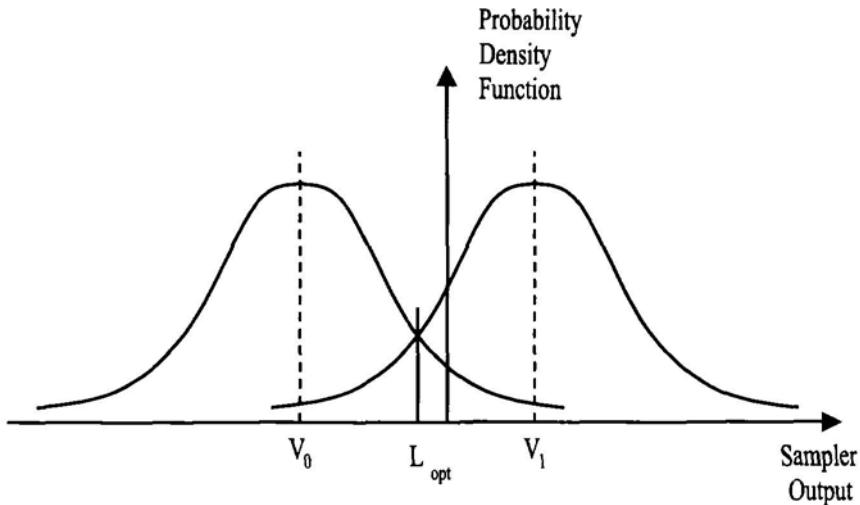


Figure 2.18. Conditional probability density function at the output of the receiver's sampler.

Now that we have found the optimum threshold level, the next step is to find the optimum filter to minimize the probability of error. The probability of error given that waveform  $s_0(t)$  was transmitted is:

$$\begin{aligned} P(e/s_0) &= \int_{\frac{V_0+V_1}{2}}^{\infty} e^{-\frac{(x-V_0)^2}{2\sigma_N^2}} \\ &= \frac{1}{2} \operatorname{erfc}\left(\frac{V_1 - V_0}{2\sqrt{2}\sigma_N}\right) \end{aligned} \quad (2.36)$$

Where,  $\operatorname{erfc}(x)$  is the complementary error function, which is defined as:

$$\begin{aligned} \operatorname{erfc}(x) &= 1 - \operatorname{erf}(x) \\ &= \frac{2}{\sqrt{\pi}} \int_x^{\infty} e^{-t^2} dt \end{aligned} \quad (2.37)$$

$\operatorname{erf}(x)$  is the error function. Another function related to the complementary error function is the Q-function, which is given by:

$$\begin{aligned} Q(x) &= \frac{1}{\sqrt{2\pi}} \int_x^{\infty} e^{-t^2/2} dt \\ &= \frac{1}{2} \operatorname{erfc}\left(\frac{x}{\sqrt{2}}\right) \end{aligned} \quad (2.38)$$

Similarly, it can be shown that the probability of error given that waveform  $s_1(t)$  was transmitted is:

$$P(e/s_1) = \frac{1}{2} \operatorname{erfc}\left(\frac{V_1 - V_0}{2\sqrt{2}\sigma_N}\right) \quad (2.39)$$

Hence, the average probability of error, at the output of the decision block, is:

$$\begin{aligned} P_e &= \sum_i P(s_i)P(e/s_i) \\ &= \frac{1}{2} \operatorname{erfc}\left(\frac{V_1 - V_0}{2\sqrt{2}\sigma_N}\right) \end{aligned} \quad (2.40)$$

The complementary error function,  $\operatorname{erfc}(x)$ , decreases monotonically with  $x$ . Hence, the optimum filter that minimizes the probability of error has a transfer function that maximizes the ratio:

$$\Delta = \frac{V_1 - V_0}{\sigma_N} \quad (2.41)$$

Maximizing  $\Delta$  is equivalent to maximizing  $\Delta^2$ . Using equations 2.28 and 2.30 we arrive at the following expression:

$$\Delta^2 = \frac{\left| \int_{-\infty}^{\infty} [S_1(f) - S_0(f)] H_R(f) e^{j2\pi f T_b} df \right|^2}{\frac{N_o}{2} \int_{-\infty}^{\infty} |H_R(f)|^2 df} \quad (2.42)$$

Apply Schwarz's inequality [14] to equation 2.42. Schwarz inequality states that:

$$\left| \int_{-\infty}^{\infty} U(f) V^*(f) df \right|^2 \leq \int_{-\infty}^{\infty} |U(f)|^2 df \int_{-\infty}^{\infty} |V(f)|^2 df \quad (2.43)$$

Where  $U(f)$  and  $V(f)$  are arbitrary complex functions. The equality holds if  $U(f) = kV(f)$ , where  $k$  is a constant. In equation 2.42 take:

$$U(f) = H_R(f) \quad (2.44)$$

and

$$V^*(f) = [S_1(f) - S_0(f)] e^{-j2\pi f T_b} \quad (2.45)$$

Substituting equation 2.42 into Schwarz's inequality we get:

$$\Delta^2 \leq \frac{2}{N_o} \int_{-\infty}^{\infty} |S_1(f) - S_0(f)|^2 df \quad (2.46)$$

The equality holds if  $U(f)$  and  $V(f)$  are linearly related. Hence, the optimum receiver filter that maximizes  $\Delta^2$  and consequently minimizes the bit error rate is:

$$H_{R,\text{opt}}(f) = [S_1^*(f) - S_0^*(f)] e^{-j2\pi f T_b} \quad (2.47)$$

The constant of proportionality  $k$ , which represents the receive filter gain was set to 1. Taking the inverse Fourier transform of equation 2.47, we get the impulse response of the optimum receive filter:

$$h_{R,\text{opt}} = s_1(T_b - t) - s_0(T_b - t) \quad (2.48)$$

The impulse response of the optimum receiver filter is matched to the difference between the transmitted waveforms  $s_0(t)$  and  $s_1(t)$ . Hence, this receiver is known as the matched filter receiver. The output of the sampler following the receiver filter can be expressed as:

$$V_i = \int_0^{T_b} s_i(t) [s_0(t) - s_1(t)] dt \quad (2.49)$$

Figure 2.19 shows a receiver implementation of equation 2.49. This receiver is known as the correlation receiver. The probability of error at the output of the sampler of Figure 2.19 is found by combining equations 2.40, 2.41 and 2.42 with the maximum value of  $\Delta^2$  from inequality 2.46:

$$P_e = \frac{1}{2} \operatorname{erfc} \left( \frac{1}{2N_o} \sqrt{\int_{-\infty}^{\infty} |S_1(f) - S_0(f)|^2 df} \right) \quad (2.50)$$

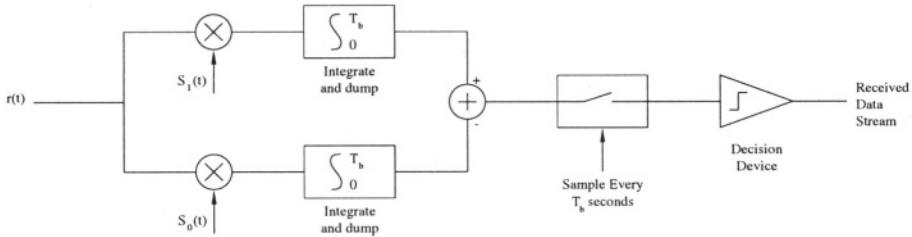


Figure 2.19. Correlation receiver.

Using Parseval's theorem:

$$\begin{aligned} \int_{-\infty}^{\infty} |S_1(f) - S_0(f)|^2 df &= \int_{-\infty}^{\infty} [s_1(t) - s_0(t)]^2 dt \\ &= E_{b1} + E_{b0} - 2\rho\sqrt{E_{b1}E_{b0}} \\ &= 2E_b \left( 1 - \frac{2\rho\sqrt{E_{b1}E_{b0}}}{E_{b1} + E_{b0}} \right) \end{aligned} \quad (2.51)$$

Where,

$E_b$  is the average bit energy:

$$E_b = \frac{E_{b1} + E_{b0}}{2} \quad (2.52)$$

$\rho$  is the correlation coefficient between waveforms  $s_0(t)$  and  $s_1(t)$ .  $\rho$  is given by:

$$\rho = \frac{1}{\sqrt{E_{b1}E_{b0}}} \int_{-\infty}^{\infty} s_1(t)s_0(t)dt \quad (2.53)$$

Where,  $-1 \leq \rho \leq 1$ .

To minimize the probability of error for a fixed average bit energy ( $E_b$ ), the argument of the complementary error function is maximized. This is equivalent to maximizing equation 2.51. For a fixed  $E_b$ , the maximum value of equation 2.51 is 2 and this occurs when:

1.  $\rho = -1$ , i.e. the two waveforms are antipodal.
2.  $E_{b_1} = E_{b_0}$ , i.e. the waveforms are of equal energy.

Hence, the minimum probability of error for a binary system using matched filters is;

$$P_{e,\text{opt}} = \frac{1}{2}\text{erfc}\sqrt{\frac{E_b}{N_o}} \quad (2.54)$$

The matched filter achieves the most optimum bit error rate performance in the presence of AWGN. However, it doesn't guarantee inter-symbol interference free transmission. To achieve inter-symbol interference free transmission, the overall frequency response of the system (the transmitter filter, the channel and the receiver filter) must satisfy the Nyquist criteria of equation 2.7. Assume that  $H_{\text{nyq}}(f)$  is a function that satisfies this condition. Hence, the transmit and receive filters are related by the following equation:

$$H_{\text{nyq}}(f) = H_T(f)H_R(f) \quad (2.55)$$

Where,

$H_T(f)$  is the transmit filter,

$H_R(f)$  is the receive filter.

The channel is assumed to have a flat spectral response in the desired bandwidth.

In order to satisfy the optimum bit error rate performance, the transmit and receive filters are matched according to the following condition:

$$H_R(f) = H_T^*(f)e^{-j2\pi fT_b} \quad (2.56)$$

Hence,

$$H_{\text{nyq}}(f)e^{j2\pi fT_b} = |H_T|^2 \quad (2.57)$$

The exponential term  $e^{j2\pi fT_b}$  is a delay term that can be arbitrary included in the frequency response of the Nyquist function  $H_{\text{nyq}}(f)$ . Furthermore, assuming that  $H_T(f)$  is real, therefore:

$$H_T(f) = \sqrt{H_{\text{Nyq}}(f)} \quad (2.58)$$

and

$$H_R(f) = \sqrt{H_{\text{Nyq}}(f)} e^{-j2\pi f T_b} \quad (2.59)$$

Dropping the exponential term, which represents a delay we get:

$$H_R(f) = \sqrt{H_{\text{Nyq}}(f)} \quad (2.60)$$

Figure 2.20 shows the transmit and receive filters for a wireless system that satisfies the Nyquist criteria for no inter-symbol interference and the matched filter criteria for optimum bit error rate performance.

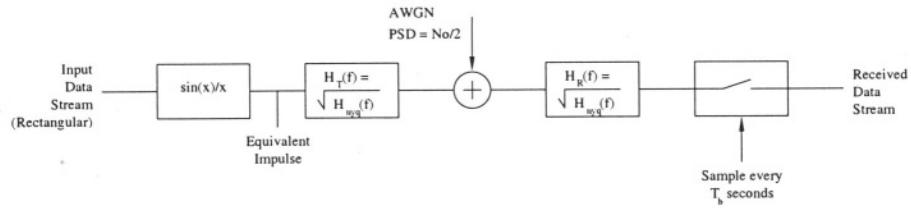


Figure 2.20. Wireless system satisfying the Nyquist criteria for no inter-symbol interference and the matched filter criteria for optimum bit error rate performance.

## Chapter 3

# THE MOBILE RADIO

### 3.1 INTRODUCTION

In this chapter, we present the characteristics of the cellular system and discuss the channel impairments that affect the design of a mobile cellular system. In section 3.2, we introduce the cellular concept which involves dividing the coverage area into small cells and reusing the same frequency in cells that are separated by a large enough distance. As the mobile terminal moves from one cell to the next, the telephone call is handed-off from one base station to the next. The handoff concept is also introduced in section 3.2. The channel impairments that influence the propagation of radio waves are presented in section 3.3. Large scale fading models that are used to determine the average received power are presented in section 3.4. The motion of the mobile terminal relative to the base station leads to a deviation in the frequency received by the mobile terminal (base station) relative to that transmitted by the base station (mobile terminal). This is known as the Doppler Shift, which is presented in section 3.5. Small scale fading manifests itself in large variations in the received signal strength over short distances. In section 3.6, we present the different types of small scale fading due to multi-path propagation and the time variant nature of the channel. Finally, in section 3.7, we talk about the diversity techniques used to enhance the performance of wireless communication systems.

### 3.2 THE CELLULAR CONCEPT

A simple mobile telephone system that covers an entire metropolitan area would consist of a high-power transmitter connected to a huge antenna located close to the center of the coverage region. However, such a system will quickly run into several problems:

1. Spectral congestion. The amount of spectrum allocated to such a system is limited. Hence, the maximum number of users is limited. Consider a mobile system with a 50 MHz spectrum and having 60 KHz full duplex channels (30 KHz for the uplink and 30 KHz for the downlink). The total number of users that can simultaneously use the system is 833. For a city with a population of 1 million people, this is less than 0.1% of its population.

The first Bell mobile system to serve New York City in the early seventies could only support 12 simultaneous calls, while it had a coverage area of 1000 square miles.

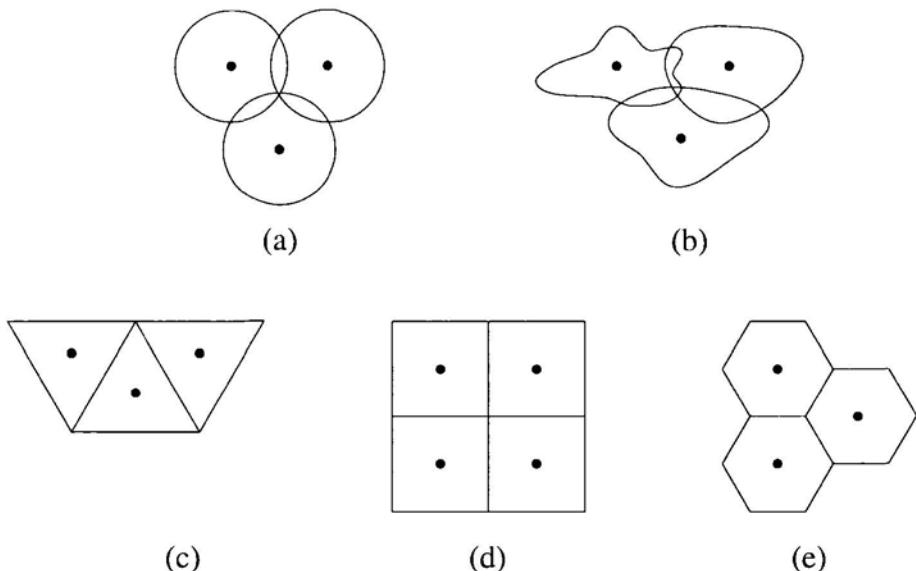
2. Having a large coverage area means that the mobile terminal would have to transmit large power to guarantee the reception of the signal by the base station when the mobile terminal is at the edge of the coverage area. This leads to shorter battery lifetime. In addition, the high transmission power level can cause health problems.
3. The near-far problem. This happens when two terminals are transmitting to the base station. One of them is close to the center of the coverage area and the other is near the edge. The received signal from the near mobile terminal can be several orders of magnitude larger than that from the far mobile terminal. This increases the receiver's linearity and selectivity requirements.

A new concept is required to be able to provide national roaming. This concept is the cellular concept. The coverage area is divided into cells. Each cell is serviced by a base station located near the center of the cell. The frequency assigned to the cellular system is divided between the cells in such a way that no two adjacent cells are allocated the same part of the spectrum. The same spectrum can be allocated (*reused*) in cells that are separated by a minimum distance to insure that interference is at an acceptable level.

### **3.2.1 Cell Shape**

Ideally, the coverage area of each base station should be a circle with the base station located at its center. However, due to the shadowing effects of buildings and irregular terrain outline (such as mountains), the shape of the coverage area becomes irregular as shown in Figure 3.1.b.

When analyzing the cellular structure, it is desirable to choose the cells in such a way that they perfectly interlock with each other without overlapping or leaving any gaps. Neither the irregular cell shape shown in Figure 3.1.b, nor the ideal circular shape shown in Figure 3.1.a satisfy this requirement. However, this requirement is satisfied by choosing the cell shape to be a regular polygon. There are three types of regular polygons that can interlock perfectly with no gaps or overlapping regions, these are the triangle, square and hexagon shown



*Figure 3.1.* Cell shape: (a) Ideal cell shape (circular). (b) Real cell shape (irregular). (c) Hypothetical cell shape (triangular). (d) Hypothetical cell shape (square). (e) Hypothetical cell shape (hexagonal).

in Figure 3.1.c-e. Of these three polygons, the hexagon resembles the circle (ideal shape) the most. Hence, when analyzing the coverage of a cellular system, each cell is assumed to have a hexagonal foot print. This is just a hypothetical model for analysis.

The base station can be located at the center of the cell or at its corner. If it is located at the center of the cell, as shown in Figure 3.2.a, an omni-directional antenna is used. However, if the base station is located at the corner of the cell, as shown in Figure 3.2.b, a 120-degree sectorial antenna is used.

### 3.2.2 Cell Clustering

The spectrum allocated to the cellular system is limited. To increase the system capacity, the same portion of spectrum is reused in different cells. However, this leads to co-channel interference. Co-channel interference refers to the phenomena where the receiver receives two or more signals, as shown in Figure 3.3, having the same frequency and during the same time slot. One of these signals is from the desired transmitter which is located in the receiver's cell. The other signals are from other transmitters located in different cells but are assigned the same portion of the spectrum.

The cells of the cellular system are divided into clusters. The cells of each cluster are assigned different portions of the frequency spectrum allocated to

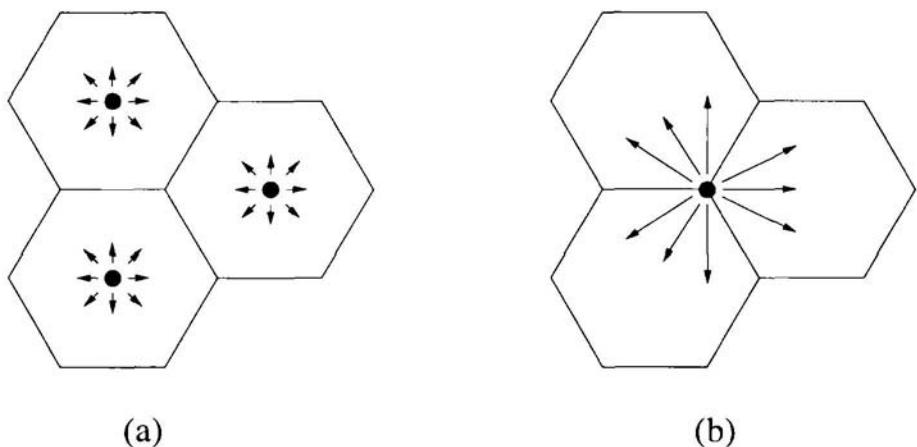


Figure 3.2. Base station location: (a) At the center of the cell. (b) At the corner of the cell.

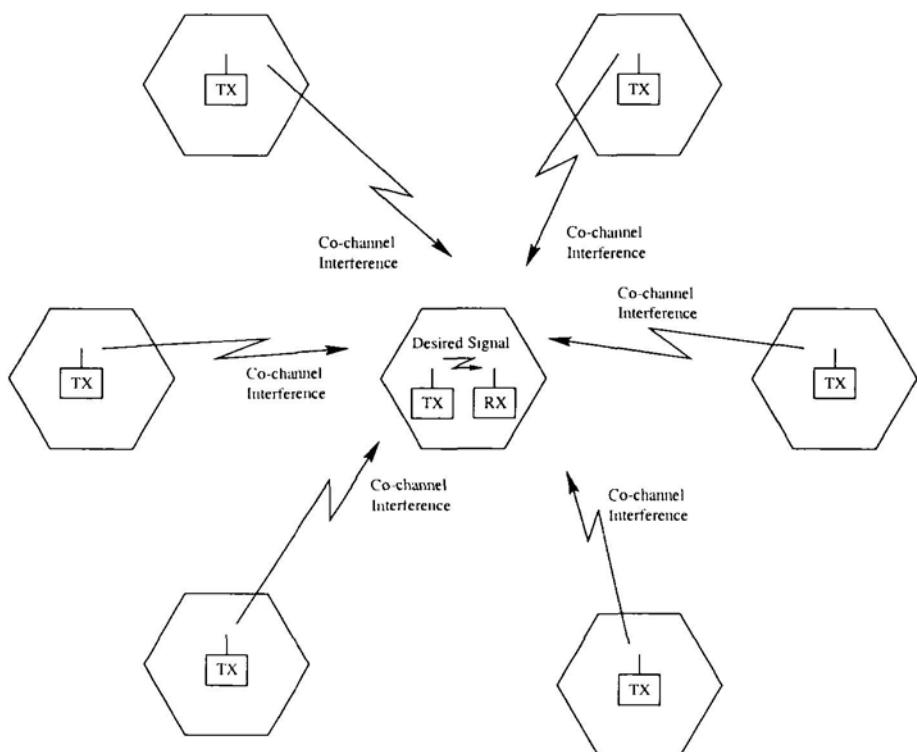


Figure 3.3. Co-channel interference in the cellular system.

the cellular system. The frequency spectrum is then reused in the other clusters. The number of cells in one cluster is known as the reuse factor. The larger the reuse factor, the larger the relative separation between two cells using the same portion of the frequency spectrum, and hence the lower the co-channel interference. However, the larger the reuse factor, the smaller the frequency band allocated to each cell, the fewer the number of radio channels, and hence the lower the system capacity.

A compromise is usually required in determining the reuse factor. The smallest reuse factor, which meets the system's co-channel interference requirement, is used in order to maximize the system's capacity.

When the hypothetical hexagonal cell shape is used the reuse factor  $N$  satisfies the following equation:

$$N = i^2 + ij + j^2 \quad (3.1)$$

Where  $i$  and  $j$  are non-negative integers. When  $i = 1$  and  $j = 0$ , the reuse factor is 1. In this case, the same spectrum is used in every cell, and the co-channel interference is highest. Figure 3.4 shows the relative positioning of two cells allocated the same portion of the frequency spectrum and having  $i = 2$  and  $j = 3$ . In practice, we use  $i = 2$  and  $j = 1$  which gives a reuse factor of 7. This is shown in Figure 3.5, where the shaded cells use the same portion of the spectrum.

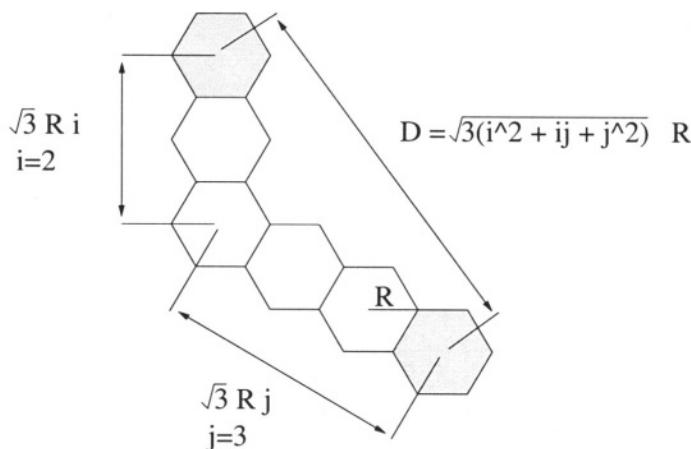
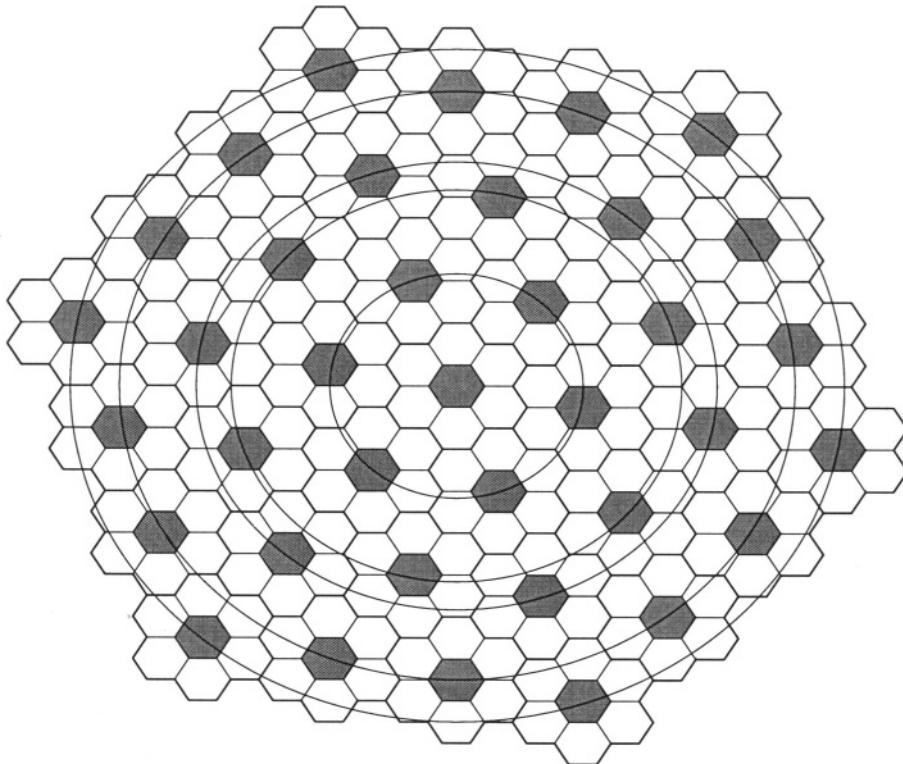


Figure 3.4. Relative positioning of two cells allocated the same portion of the frequency spectrum, and having  $i = 2$ , and  $j = 3$ .  $D$  is the reuse distance.

The reuse distance ( $D$ ) is defined as the minimum distance between the center of two cells allocated the same portion of the spectrum. According to Figure 3.4,  $D$  is related to  $i$  and  $j$  of equation 3.1 by the following equation:



*Figure 3.5.* Cellular system cell plan having a cluster size of seven cells.

$$\begin{aligned} D &= \sqrt{3(i^2 + ij + j^2)}R \\ &= \sqrt{3N}R \end{aligned} \quad (3.2)$$

For a reuse factor of seven, the reuse distance is  $\sqrt{21}R$ . As shown in Figure 3.5, there are six cells using the same portion of the frequency spectrum, that are at a distance of  $\sqrt{21}R$ . The next tier of cells that utilize the same portion of the frequency spectrum are at a distance of  $\sqrt{63}R$ , there are also six cells in this tier. The following tier of six cells utilizing the same portion of the frequency spectrum is at a distance of  $\sqrt{84}R$ , and so on.

The cells of the first tier are the most significant contributors to the co-channel interference. The effect of the co-channel interference produced by the cells of the second and higher order tiers can be neglected, as illustrated in the following example.

**Example 3.1.** For a cellular system having a reuse factor of 7, calculate the ratio between the co-channel interference produced by one cell of the first tier

to that produced by one cell of the second tier. Assuming that the decay of the received power ( $P_r$ ) with distance ( $d$ ) is given by:

- (a)  $P_r \propto \frac{1}{d^2}$ .
- (b)  $P_r \propto \frac{1}{d^4}$ .

**Solution.** (a) Assume that the received power is given by:

$$P_r = \frac{k}{d^2}$$

Where,  $k$  is a constant.

The co-channel interference due to a first tier cell is:

$$P_{r1} = \frac{k}{21R^2}$$

The co-channel interference due to a second tier cell is:

$$P_{r2} = \frac{k}{63R^2}$$

Therefore,

$$\begin{aligned} \frac{P_{r1}}{P_{r2}} &= 3 \\ &= 4.8 \text{dB} \end{aligned}$$

(b) Assume that the received power is given by:

$$P_r = \frac{k}{d^4}$$

The co-channel interference due to a first tier cell is:

$$P_{r1} = \frac{k}{(21R^2)^2}$$

The co-channel interference due to a second tier cell is:

$$P_{r2} = \frac{k}{(63R^2)^2}$$

Therefore,

$$\begin{aligned} \frac{P_{r1}}{P_{r2}} &= 9 \\ &= 9.5 \text{dB} \end{aligned}$$

The results of this example indicate that when the propagation model follows that of free space, i.e. the received power decreases with the square of the distance, the co-channel interference caused by a second tier cell is one third that caused by a first tier cell. In this case, neglecting the co-channel interference produced by a second tier cell might not be an accurate approximation.

However, when the received power decreases with the fourth power of the distance, the co-channel interference produced by a second tier cell is about 11% of that produced by a first tier cell. In this case, neglecting the co-channel interference produced by higher tier cells is a more reasonable approximation.

**Example 3.2.** For a cellular system having a reuse factor of 7, calculate the carrier to co-channel interference (C/I) ratio for a mobile terminal located at the edge of the cell. Neglect the co-channel interference produced by second and higher tier cells. Assume that the decay of the received power ( $P_r$ ) with distance ( $d$ ) is given by:

- (a)  $P_r \propto \frac{1}{d^2}$ .
- (b)  $P_r \propto \frac{1}{d^4}$ .

**Solution.** A mobile terminal located at the edge of the cell is at a distance  $R$  from the base station located at the center of the cell. Therefore, the received power  $C$  is given by:

$$C = \frac{k}{R^n}$$

Where,  $n$  is 2 or 4 depending on the propagation model.

The interference is produced by an interferer located in a first tier cell, which is at a distance  $D$ .  $D$  is the reuse distance. The co-channel interference produced by one first tier cell is:

$$i = \frac{k}{D^n}$$

Since there are 6 first tier interferes, the total co-channel interference produced by the first tier cells is:

$$I = 6i = \frac{6k}{D^n}$$

Therefore,

$$C/I = \frac{D^n}{6R^n}$$

For a cellular system having a reuse factor of seven,

$$D = \sqrt{21}R$$

Therefore,

$$\begin{aligned} C/I &= \frac{(21)^{n/2}}{6} \\ &= 5.44 \text{ dB} \quad \text{for } n = 2 \\ &= 18.66 \text{ dB} \quad \text{for } n = 4 \end{aligned}$$

### 3.2.3 Cell Splitting

As the number of users starts to increase in a certain region, e.g. metropolitan areas, to a point where there isn't enough radio channels to support the desired grade of service, the cells of that region are subdivided into smaller cells. Hence, creating more radio channels and increasing the system capacity.

Practically, the cells are quite small close to urban centers and get larger as we go into suburban and rural areas. A cellular system employing the cell splitting concept is shown in Figure 3.6.

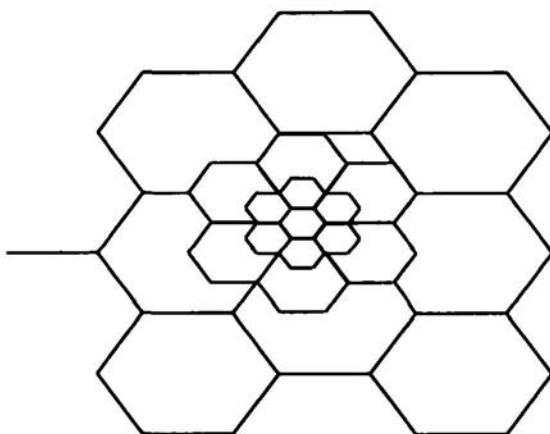


Figure 3.6. Cell splitting is used to increase the capacity of the cellular system in metropolitan areas.

### 3.2.4 Handoff

Handoff, which is also known as handover in European wireless systems, is the process of transferring a mobile terminal from one base station to the next as the user roams from one cell to another. This process is transparent and imperceptible to the user. The handoff process is complete when a new radio channel is allocated to the mobile terminal in the new cell, and the old radio channel of the old cell is relinquished.

As the mobile terminal moves towards the edge of the coverage area of a particular base station, the reception becomes weak. There is a minimum usable received signal strength, for acceptable voice quality in a cellular system. In a noise limited environment this signal strength is -100 dBm, while in an interference limited environment it is -95 dBm [15]. The handoff threshold (the received signal level at which handoff is initiated) is slightly larger than the minimum usable received signal strength.  $\Delta P$  is the difference between these two power levels. Hence,

$$\Delta P = P_{\text{Handoff}} - P_{\text{Min usable}} \quad (3.3)$$

The choice of  $\Delta P$  is a compromise between the volume of handoffs, which increases as  $\Delta P$  becomes smaller, and the allowed handoff delay, which increases as  $\Delta P$  becomes larger. The handoff delay is the time between the initiation of the handoff process, when the received signal power falls below the handoff threshold power level, and the completion of the handoff process. If the handoff delay is too large, the call might be dropped before the handoff process is complete.

In first generation analog cellular systems, the base station monitors the received signal strength indicator (RSSI), this information is passed on to the MSC where it is processed to decide if a handoff is required or not. The typical handoff delay time for first generation systems is 10 seconds, and  $\Delta P$  is in the order of 6 dB to 12 dB.

The type of handoff control used in first generation systems is centralized handoff control, as the number of mobile terminals increases and the cell size decreases, this type of handoff becomes inefficient in handling the increase in demand [16]. Second generation cellular systems use microcells [17] in urban areas, which have an area two orders of magnitude less than that of standard cells, this leads to a corresponding increase in the capacity of the system, at the same time, the volume of handoffs increases. To support these loads efficiently, handoff decisions have to be mobile assisted.

In second-generation digital cellular systems, the mobile terminals monitor the signal strength from the surrounding base stations, and passes on these measurements to the serving base stations. The handoff process is initiated when the power received from a neighboring base station exceeds that of

the serving base station by a certain level or for a certain period of time. The handoff delay of second-generation systems is 1 to 2 seconds, this is considerably shorter than that of first generation systems.

It is important to complete the handoff process successfully to avoid the annoyance of abruptly terminating a telephone conversation. There are certain circumstances which occur when a handoff is necessary, yet it can't be made. For example, if the received signal strength falls below the handoff threshold when the mobile terminal is located within the cell and far away from its boundary. This can occur due to fading or shadowing effects. Another example is when the mobile terminal moves to a new cell where there is no available radio channels. In these cases, the call is abruptly terminated.

From the users point of view, abruptly terminating a telephone conversation is more annoying than occasionally blocking a new one. Hence, handoff requests are prioritized over call initiation requests. For example, a group of channels, known as the guard channels, can be reserved to handle handoff requests. This method prioritizes handoffs, yet it reduces the overall system capacity [18]. Another method is to queue the blocked handoffs, instead of terminating the call abruptly [15], [18].

In the CDMA digital cellular system, the mobile terminal can establish a radio link between two or more base stations as it approaches the boundary of a cell. In the CDMA system, all the cells use the same radio frequency. Hence, all the radio links established have same frequency, however, each radio link uses a different pseudonoise (PN) code. The mobile switching center evaluates the quality of the received signal at the different base stations, and uses the one with the highest quality. As the mobile terminal moves towards the center of the new cell, the other radio links are relinquished. This type of handoff is called soft handoff. Unlike hard handoff, soft handoff can support several radio links to different base stations simultaneously because, all base stations operate at the same frequency. In a hard handoff scenario, each base station is allocated a different frequency band, as the mobile terminal migrates from one cell to the next, it must switch to the frequency of the new radio link established in the new cell, and drop that of the old cell.

Occasionally, as the user roams around, the mobile terminal crosses the boundary between two regions served by two different mobile switching centers (MSCs). In this case, an intersystem handoff must take place. Consider the scenario of Figure 3.7, initially, the call originates in region A which is served by MSCA, eventually, the user reaches a cell at the boundary of the service area of MSCA, and begins to move out of that service area. MSCA attempts to find a new base station within its service area to handoff the mobile terminal to, but fails to do so. MSCA sends an intersystem handoff request to MSCB, which intervenes to complete the handoff process by assigning the mobile terminal to a base station in its service area.

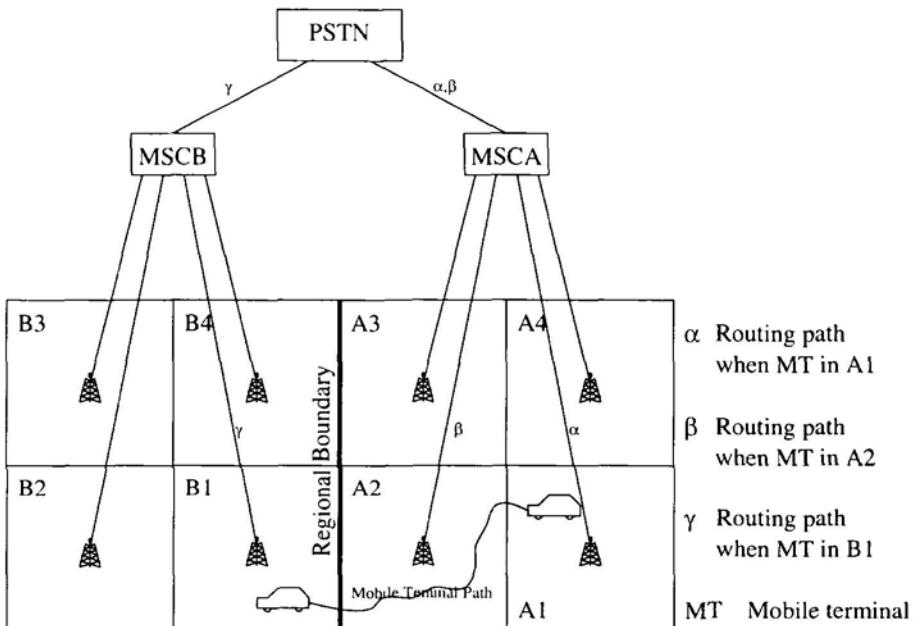


Figure 3.7. Intersystem handoff.

There are certain issues which need to be considered when implementing an intersystem handoff. For example, the two mobile switching centers might be from two different vendors, hence intersystem compatibility becomes an issue which must be addressed. Furthermore, when migrating from one region to another, the mobile terminal might be leaving or entering its home region and a local call might become a long distance call or vice versa. Hence, long distance and roaming charges need to be billed correctly and agreed upon before an intersystem handoff can occur.

### 3.3 CHANNEL IMPAIRMENTS

Information is conveyed between the base station and the mobile terminal of a wireless communications system by radio waves. Radio waves are a subset of electromagnetic waves. The radio waves employed in the cellular and PCS systems have a frequency range of 400 MHz to 2.5 GHz, which is equivalent to a wavelength in the range of 12 cm to 75 cm.

As the radio wave propagates from the transmitter to the receiver, several detrimental factors start influencing it, which impact the overall system performance:

1. Propagation path loss with distance.

2. Shadowing effects.
3. Multi-path fading.
4. Doppler effect.
5. Noise.

In this section, we define each of these factors and determine its effect on the reception of radio waves. For a more detailed analysis of these effects and the models, which describe them, the reader is referred to [19] and [18].

Propagation models that determine the average received power for an arbitrary location are called large-scale propagation models. Whereas, propagation models that determine the rapid fluctuations of the received power over short distances, comparable to the wavelength, are called small-scale fading models.

The electromagnetic wave consists of two mutually orthogonal fields, the electric field and the magnetic field, these fields are also orthogonal to the direction of propagation. The electric field (as well as the magnetic field) are phasor quantities, they have magnitude and phase. The magnitude of the electric field ( $|E|$ ) is related to the power density by the following equation:

$$S = \frac{|E|^2}{Z_o} \quad (3.4)$$

Where,  $Z_o$  is the intrinsic impedance of free space, which equals  $120\pi$  or 377 Ohms.

### 3.3.1 Free Space Propagation

The free space propagation model is used when there is an unobstructed line-of-sight path between the receiver and the transmitter and the region between them is free from any objects that absorb or reflect the RF electromagnetic wave. Furthermore, the atmosphere must behave as a non-absorbing material, and the reflection off the surface of the earth can be neglected. In this case, the received power is determined by the inverse-square law.

Assume that the power radiated by an isotropic source is  $P_T$ . An isotropic source is a source that radiates equally in all directions. The power density at a distance  $d$  from that isotropic transmitter is given by:

$$S_R = \frac{P_T}{4\pi d^2} \quad (3.5)$$

The antenna used in the transmitter usually has some directivity, it transmits more power in one direction than in another. This fact implies directional gain ( $G$ ). The directional gain is defined as the ratio between the input power to the antenna whose gain is being calculated, to the input power of an isotropic

antenna, such that the maximum power density at a distance  $d$  is equal for the two antennas.

The product  $PG$  is known as the effective isotropic radiated power (EIRP).  $P$  is the input power to an antenna whose directional gain is  $G$ .

Assuming that the transmitter antenna has a directional gain  $G_T$ , the maximum power density at a distance  $d$  from this transmitter is given by:

$$S_R = \frac{G_T P_T}{4\pi d^2} \quad (3.6)$$

The received power depends on the aperture area of the receiver antenna. It is given by:

$$P_R = S_R A_R \quad (3.7)$$

$A_R$  is the aperture area of the receiver antenna.  $A_R$  is related to the receiver antenna directional gain ( $G_R$ ) by the following equation:

$$A_R = \frac{G_R \lambda^2}{4\pi} \quad (3.8)$$

Combining equations 3.6, 3.7 and 3.8, the received power becomes:

$$P_R = \frac{G_T G_R \lambda^2}{(4\pi)^2 d^2} P_T \quad (3.9)$$

Equation 3.9 is known as Friis free space equation [18]. Expressed in decibels (dBs) it becomes:

$$\begin{aligned} P_R(\text{dBm}) &= P_T(\text{dBm}) + G_T(\text{dB}) + G_R(\text{dB}) \\ &\quad - 20 \log(F_{\text{MHz}}) - 20 \log(d) + 27.56 \end{aligned} \quad (3.10)$$

According to equation 3.10, the received power decays with distance at a rate of 20 dB/decade.

**Example 3.3.** An antenna transmits a 1-watt signal having a carrier frequency of 900 MHz. Calculate the received power in dBm at a distance of 1 km from the transmitter. Assume that both the receiver and transmitter antennas are halfwave dipoles with a gain of 1.65.

**Solution.**

$$\text{Transmitter Power} = 10 \log(1000 \text{ mW}) = 30 \text{ dBm}$$

$$G_R = G_T = 10 \log(1.65) = 2.17 \text{ dB}$$

Therefore,

$$\begin{aligned} P_R &= 30 + 2.17 + 2.17 - 20 \log(900) - 20 \log(1000) + 27.56 \\ &= -57.18 \text{ dBm} \end{aligned}$$

### 3.3.2 The Two-Ray Propagation Model

In a wireless communications system, reflected rays usually exist between the base station and the mobile terminal in addition to the direct line of sight (LOS) rays. Hence, the free space propagation model fails to describe the path attenuation with distance in this case.

The reflected ray (or rays) can be reflected off the ground or man made constructions. In this section, we drive the propagation model for a two ray wireless channel. There is the direct LOS ray and the ground reflected ray.

Whenever there is a discontinuity between two media, and an electromagnetic wave is incident on this discontinuity, the wave is partly reflected and partly transmitted at the boundary. The ratio between the reflected wave and the incident wave, as well as the ratio between the transmitted wave and the incident wave is determined by Snell's law [20]. These ratios depend on the polarization of the electromagnetic wave (the direction of the electric field), and on the angle of incidence (the angle between the incident ray and the normal to the boundary).

When the angle of incidence  $\theta_i$  approaches  $90^\circ$ , the reflected wave is equal in magnitude and  $180^\circ$  out of phase with the incident wave, i.e. the reflection coefficient  $\Gamma = -1$ .

Figure 3.8 shows a transmitter having a height  $h_T$ , and a receiver having a height  $h_R$ , and separated by a distance  $d$ . Two rays propagate from the transmitter to the receiver. The direct LOS ray and the ground reflected ray. The direct LOS ray is of length:

$$d_l = \sqrt{(h_T - h_R)^2 + d^2} \quad (3.11)$$

Assuming  $d \gg h_T, h_R$ . Therefore,

$$d_l \simeq d \left[ 1 + \frac{(h_T - h_R)^2}{2d^2} \right] \quad (3.12)$$

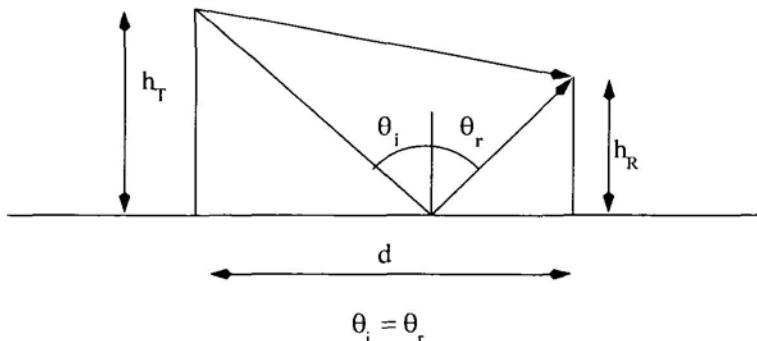


Figure 3.8. Two ray propagation model: Direct line of sight (LOS) ray and ground reflected ray.

The ground reflected ray is of length:

$$d_r = \sqrt{(h_T + h_R)^2 + d^2} \quad (3.13)$$

Again, assuming  $d \gg h_T, h_R$ . Therefore,

$$d_r \simeq d \left[ 1 + \frac{(h_T + h_R)^2}{2d^2} \right] \quad (3.14)$$

The path difference  $\Delta d = d_r - d_l$  is given by:

$$\Delta d = \frac{2h_T h_R}{d} \quad (3.15)$$

This path difference leads to a phase difference between the reflected and direct LOS rays, which is given by:

$$\begin{aligned} \Delta\phi &= \frac{2\pi}{\lambda} \Delta d \\ &= \frac{4\pi h_T h_R}{\lambda d} \end{aligned} \quad (3.16)$$

In addition, the ground reflected wave undergoes a phase shift of  $180^\circ$  at the boundary. Since  $d \gg h_T, h_R$ , therefore:

$$d_l \simeq d_r \simeq d \quad (3.17)$$

Hence, according to equations 3.4 and 3.6:

$$|E_l| = |E_r| = \frac{\sqrt{P_T G_T Z_o}}{\sqrt{4\pi d}} \quad (3.18)$$

Where,

$E_l$  is the electric field of the direct LOS ray at the receiver.

$E_r$  is the electric field of the ground reflected ray at the receiver.

The total electric field at the receiver  $E_t$  is given by:

$$|E_t| = |E_l| |1 - e^{j\Delta\phi}| \quad (3.19)$$

The minus sign is because the reflected wave undergoes a  $180^\circ$  phase shift at the boundary. Therefore:

$$|E_t| = 2|E_l| \sin\left(\frac{\Delta\phi}{2}\right) \quad (3.20)$$

If  $\Delta\phi \ll \pi$ , therefore according to equation 3.16:

$$|E_t| \simeq |E_l| \frac{4\pi h_T h_R}{\lambda d} \quad (3.21)$$

Combining equations 3.4, 3.7, 3.18 and 3.21, we get the following expression for the total power ( $P_R$ ) at the receiver:

$$P_R = \frac{G_T G_R h_T^2 h_R^2}{d^4} P_T \quad (3.22)$$

According to equation 3.22, when  $d \gg h_T, h_R$ , the received power decreases with the fourth power of the distance or at a rate of  $40dB/\text{decade}$ . Note also that the path loss is independent of frequency. The two ray model has been found to be an accurate model for predicting the value of the signal at a large distance from the transmitter (several kilometers).

### 3.4 LARGE SCALE FADING

The large scale fading models determine the average signal power at a certain location. It takes account of two phenomena. First, the path attenuation with distance, this determines the mean value of the received signal at a certain distance from the transmitter, it is described in terms of the  $\nu$ -power law. Second, the shadowing effect of the surrounding environment, this is the variation of a signal power about the mean value determined by the  $\nu$ -power law, this variation is described by a log-normal distribution.

The theoretical analysis presented in the previous section indicates that the received signal power decreases with the square of the distance in free space and with the fourth power of the distance in the case of a two ray model.

Measured results [21] indicate that the average received signal power decreases with the  $\nu$ th power of distance:

$$P_R(d) = P_R(d_o) \left( \frac{d_o}{d} \right)^\nu \quad (3.23)$$

Where,

$d_o$  is the reference distance. This is a point located in the far field of the antenna.  $d_o$  is typically 1 Km for large cells, 100 m for micro-cells, and 1 m for indoor wireless systems [22].

$\nu$  is the path loss exponent,  $\nu$  can vary between 1.6 to 6.2.

Table 3.1 [18] gives the path loss exponent for different environments.

*Table 3.1. Path loss exponent for different wireless environments.*

Environment	Path loss exponent
Urban streets (guided wave phenomena) [22]	< 2
In building LOS	1.6-1.8
Free space	2
Urban area cellular radio	2.7-3.5
Shadowed urban cellular radio	3-5
In and around suburban homes [21]	3-6.2
Obstructed in building	4-6
Obstructed in factories	2-3

Equation 3.23 predicts the average received power at a distance  $d$  from the transmitter. However, it doesn't take account of the surrounding environment. Irregular terrain variations, man-made structures, and foliage in the surrounding environment cause the received signal power to depart from the value predicted by equation 3.23 due to shadowing effects. It was found [23] that the received signal power follows a log-normal distribution:

$$P_R(d) = P_R(d_o) - \nu \log \left( \frac{d}{d_o} \right) + N_\sigma \quad (3.24)$$

Where,  $N_\sigma$  is a zero mean Gaussian distributed random variable having a standard deviation  $\sigma$ .  $N_\sigma$  and  $\sigma$  are both expressed in dBs.

### 3.5 DOPPLER SHIFT

The motion of the mobile terminal relative to the base station and the surrounding environment leads to a frequency shift between the transmitted frequency and the received one, which depends on the speed and direction of the

mobile terminal relative to the base station and the surrounding environment. Consider a mobile terminal moving as illustrated in Figure 3.9, the Doppler shift experienced by the mobile terminal is given by:

$$f_d = f_r - f_t = \frac{v}{\lambda} \cos \theta \quad (3.25)$$

Where,

$f_t$  is the frequency of the transmitted sinusoidal signal,

$f_r$  is the frequency of the received sinusoidal signal.

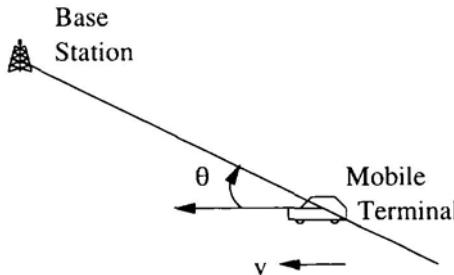


Figure 3.9. Illustration of the Doppler shift.

According to equation 3.25, the Doppler shift depends on the direction of motion of the mobile terminal relative to the base station. When the mobile terminal moves towards the base station, the Doppler shift is positive and is equal to  $v/\lambda$ . When the mobile terminal moves away from the base station, the Doppler shift is negative and is equal to  $-v/\lambda$ . When the mobile terminal moves perpendicular to the line joining it to the base station, the received signal experiences no Doppler shift.

The frequency of the received wave at the mobile terminal not only depends on the frequency of the transmitted signal, which is assumed to be a signal having a frequency  $f_t$ , but also depends on the velocity of the mobile terminal as well as the angle of arrival of the signal. The presence of multi-path components, which arrive from different directions, and hence experience different amounts of Doppler shifts, spreads the spectrum of the received signal to a frequency band centered around  $f_t$  instead of being a single spectral component. This is known as the Doppler spread.

Assume that the number of received waves at the mobile terminal is infinite. Furthermore, assume a uniform distribution for the multi-path signals arriving at all angles throughout the range  $[0, 2\pi]$ . Assume that  $P_R$  is the total received power. Hence, the power received when the angle of arrival is  $\alpha$  with an angular spread  $d\alpha$  is:

$$P_R(\alpha) = \frac{P_R}{2\pi} d\alpha \quad (3.26)$$

A received signal having an angle of arrival  $\alpha$  has a frequency given by:

$$f_\alpha = f_c + f_m \cos \alpha \quad (3.27)$$

Where,  $f_m$  is the maximum Doppler frequency. Hence,  $df$  and  $d\alpha$  are related by:

$$|df| = 2f_m \sin \alpha |d\alpha| \quad (3.28)$$

Alternatively:

$$|d\alpha| = \frac{1}{2\sqrt{f_m^2 - (f - f_c)^2}} df \quad (3.29)$$

Where  $|f - f_c| \leq f_m$ . The reason the factor 2 appeared in equation 3.29 is because  $\cos \alpha$  is an even function. Hence, a signal arriving at an angle  $\alpha$  or an angle  $-\alpha$  has the same Doppler Shift. Substituting in equation 3.26, we get:

$$P_R(f) = \frac{P_R}{4\pi \sqrt{f_m^2 - (f - f_c)^2}} df \quad (3.30)$$

Hence, the power spectrum density of the received signal is given by [24]:

$$S_R(f) = \begin{cases} \frac{P_R}{4\pi \sqrt{f_m^2 - (f - f_c)^2}} & -f_m < f - f_c < f_m \\ 0 & \text{otherwise} \end{cases} \quad (3.31)$$

Where,  $f_m$  is the maximum Doppler shift, which is given by:

$$f_m = \frac{v}{\lambda} \quad (3.32)$$

**Example 3.4.** A base station transmits a signal at 950 MHz, the signal is received by a mobile terminal moving at a speed of 100 Km/hr, and with  $\theta = 45^\circ$ ,  $\theta$  is as shown in Figure 3.9. Find the Doppler shift in the received signal.

**Solution.** The wavelength of the transmitted signal is:

$$\lambda = \frac{c}{f} = \frac{3 \times 10^8}{950 \times 10^6} = 0.316 \text{ metres}$$

The velocity of the mobile terminal is:

$$v = 100 \times \frac{1000}{3600} = 27.78 \text{ m/sec}$$

Using equation 3.25, the Doppler shift is:

$$f_d = \frac{v}{\lambda} \cos \theta = \frac{27.78}{0.316} \cos(45^\circ) = 62.2 \text{ Hz}$$

### 3.6 SMALL SCALE FADING

Small scale fading manifests itself in:

- The large variation in the envelope of the received signal over short distances, due to the constructive and destructive interference of the multi-path components.
- Time dispersion of the received signal, due to the presence of multi-path components that arrive at the receiver at different delays.
- The time variant nature of the channel, this is due to the motion of the mobile terminal and/or the motion of objects in the surrounding environment. This gives rise to Doppler frequency shift.

Figure 3.10 [18] illustrates the different types of small scale fading that can occur in a multi-path, time-variant channel. The presence of multi-path components leads to time dispersion, which gives rise to frequency selective fading or flat fading. On the other hand, the time-variant nature of the channel leads to frequency dispersion which gives rise to fast fading or slow fading

The type of fading due to the presence of multi-path components depends on the relationship between the symbol period,  $T_s$  and the maximum excess delay,  $T_m$ , which is the delay between the arrival time of the first multi-path component and that of the last multi-path component. There are two types of fading that can occur due to the time-spreading effect of multi-path propagation [22]:

1. **Frequency selective fading.** This occurs when  $T_m > T_s$ . This type of fading gives rise to pulse distortion and channel-induced intersymbol interference. It is sometimes possible to equalize this type of fading because the multi-path components are resolvable at the receiver.

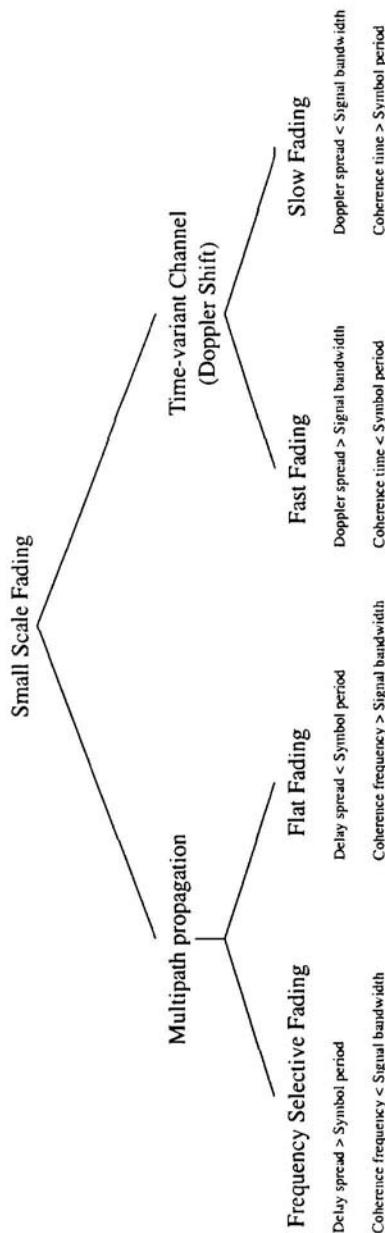


Figure 3.10. Small scale fading in a multi-path time-variant channel.

- 2. Flat fading.** This occurs when  $T_m < T_s$ . In this case, the time spread between the first and last multi-path components is smaller than the symbol duration, so there is no channel induced intersymbol interference. However, the multi-path signals can add destructively, so as to drastically reduce the level of the received signal. This leads to higher bit-error-rate.

Another way to characterize the multi-path propagation model of the wireless channel is to analyze it in the frequency domain. The coherence bandwidth,  $f_{coh}$ , is defined as the bandwidth over which two signals transmitted across the channel, have strong correlation.  $f_{coh}$  is inversely proportional to  $T_m$ , i.e. the larger the value of  $T_m$ , the lower the value of  $f_{coh}$  and vice versa.

Frequency selective fading occurs when the coherence bandwidth is smaller than the signal bandwidth, BW, i.e.  $f_{coh} < \text{BW}$ . Hence, the spectral components of the signal are not equally affected by the multi-path channel. On the other hand, flat fading occurs whenever  $f_{coh} > \text{BW}$ . Hence, all the spectral components of the signal are equally affected by the channel, accordingly, the signal shape will not be distorted, but degradation in performance occurs due to fading caused by destructive interference.

In flat fading, the probability density function of the received signal depends on the presence or absence of a non-fading components. In the former case, the received signal follows a Ricean distribution, while in the latter case the received signal follows a Rayleigh distribution. Both distributions are presented in this section.

The type of fading due to the time-variant nature of the channel depends on the relationship between the signal bandwidth, BW, which is nominally taken to be equal to the symbol rate,  $R_s$ , and the maximum Doppler shift,  $f_m$  which is proportional to the speed of the mobile terminal as given by equation 3.32. There are two types of fading that can occur due to the frequency-spreading effect of a time-variant channel [22]:

- 1. Fast fading.** This occurs when  $f_m > \text{BW}$ . In this case, the spreading of the signal in the frequency domain, due to the Doppler shift, is greater than the signal bandwidth. Hence, this type of fading gives rise to pulse distortion. A detrimental effect of fast fading is the failure of the phase locked loop (PLL) in maintaining synchronization.
- 2. Slow fading.** This occurs when  $f_m < \text{BW}$ . Hence, the frequency spreading, due to the Doppler shift, doesn't result in a significant change in the frequency spectrum of the signal. Degradation in performance occurs because there are periods of times, larger than the symbol time, during which the signal is drastically faded, this leads to higher bit-error-rate.

Another way to characterize the time-variant nature of the channel is to analyze it in the time domain. The coherence time,  $T_{coh}$ , is defined as the time

interval over which two sinusoidal signals transmitted across the channel have strong coherence. In other words, the coherence time is duration over which the channel response is essentially invariant.  $T_{coh}$  is inversely proportional to  $f_m$ , i.e. the larger the value of  $f_m$ , the lower the value of  $T_{coh}$  and vice versa.

Fast fading occurs when  $T_{coh} < T_s$ , i.e. the time during which the channel can be considered invariant is smaller than the symbol duration. Hence, the characteristics of the fading channel change several times during one symbol period, this is why it is called fast fading. Accordingly, the received symbol has a distorted pulse shape.

Slow fading occurs when  $T_{coh} > T_s$ , i.e. the time during which the channel can be considered invariant is larger than the symbol duration. Hence, the characteristics of the fading channel change slowly from one symbol to the next, this is why it is called slow fading. Accordingly, there is little distortion in the pulse shape of the received symbol.

### 3.6.1 Rayleigh Fading

In a wireless mobile channel, the received signal is made up of a number of waves with random amplitudes, angle of arrivals and phases. Because of the interference between the received waves, the resultant received signal undergoes fading. Assume that an unmodulated wave is being transmitted:

$$T(t) = \cos w_c t \quad (3.33)$$

At the antenna of the mobile terminal there are  $N$  received waves, as shown in Figure 3.11. Each wave has an amplitude  $C_n$ , a phase  $\phi_n$ , and an angle of arrival (relative to the direction of motion of the mobile terminal)  $\alpha_n$ . Associate with each angle of arrival is an angular frequency  $w_n$  which depends on the Doppler shift and is given by:

$$w_n = w_c \left( 1 + \frac{v}{c} \cos \alpha_n \right) \quad (3.34)$$

Where,  $v$  is the speed of the mobile terminal and  $c$  is the speed of light. The resultant received signal can be expressed as:

$$R(t) = \sum_{n=1}^N C_n \cos \left( w_c t + \frac{v}{c} \cos(\alpha_n) w_c t + \phi_n \right) \quad (3.35)$$

$R(t)$  can be decomposed into an in-phase term and a quadrature-phase term [25]:

$$R(t) = R_c(t) \cos(w_c t) - R_s \sin(w_c t) \quad (3.36)$$

Where,

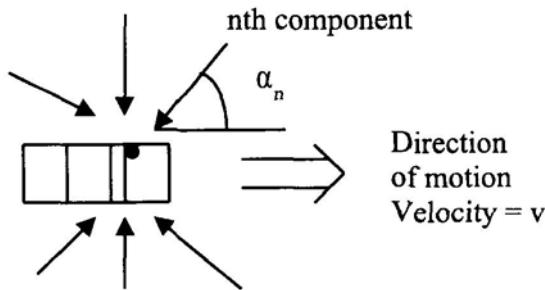


Figure 3.11. Rayleigh fading: Multi-wave reception.

$$R_c(t) = \sum_{n=1}^N C_n \cos \left( \frac{v}{c} \cos(\alpha_n) w_c t + \phi_n \right) \quad (3.37)$$

and,

$$R_s(t) = \sum_{n=1}^N C_n \sin \left( \frac{v}{c} \cos(\alpha_n) w_c t + \phi_n \right) \quad (3.38)$$

As the number of received waves  $N$  approaches infinity ( $N \rightarrow \infty$ ), and assuming that the received waves are identical and independently distributed (iid) random variables at a particular instance of time, hence according to the central limit theorem  $R_c(t)$  and  $R_s(t)$  are stationary Gaussian and independent random processes. The envelope of these two quadrature components has a Rayleigh distribution. Hence, the name Rayleigh fading.

$$E = \sqrt{R_c^2 + R_s^2} \quad (3.39)$$

Where,

$E$  is a Rayleigh distributed random variable representing the envelope at a particular time.

$R_c$  and  $R_s$  are Gaussian distributed random variables representing the amplitudes of the in-phase and quadrature-phase components at a particular instance of time.

The probability density function (pdf) of a Rayleigh distributed random variable is given by:

$$p(x) = \begin{cases} \frac{x}{\sigma^2} e^{-x^2/2\sigma^2} & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.40)$$

Where,

$\sigma$  is the root mean square (RMS) value of the received signal.

The cumulative distribution function of a Rayleigh distributed random variable is given by:

$$\text{Prob}(x \leq r) = \begin{cases} 1 - e^{-r^2/2\sigma^2} & r \geq 0 \\ 0 & r < 0 \end{cases} \quad (3.41)$$

The mean of the Rayleigh distribution is given by:

$$X_{\text{mean}} = \sigma \sqrt{\frac{\pi}{2}} \quad (3.42)$$

The variance of the Rayleigh distribution is given by:

$$X_{\sigma^2} = \sigma^2 \left( 2 - \frac{\pi}{2} \right) \quad (3.43)$$

The median of the Rayleigh distribution is given by [18]:

$$X_{\text{median}} = 1.177\sigma \quad (3.44)$$

Figure 3.12 shows the pdf and CDF of a Rayleigh distributed random variable. Figure 3.13, shows the variation of the envelop of a Rayleigh faded signal versus time.

Two important parameters of a Rayleigh fading channel are the level crossing rate (LCR) and the average fade duration, these parameters determine the type of diversity scheme and error correction code employed in the wireless communications system to combat fading. The level crossing rate is defined as the expected number of times per second the received signal envelope crosses a specified level  $R$ , in a positive going direction. The average number of level crossings per second is given by:

$$N_R = \sqrt{2\pi} f_m \rho e^{-\rho^2} \quad (3.45)$$

Where,

$f_m$  is the maximum Doppler frequency shift as given by equation 3.32.  $f_m$  increases as the speed of the mobile terminal increases.

$\rho = R/R_{\text{rms}}$  is the specified level  $R$  normalized to the local root mean square amplitude.

The average fade duration is the average period of time the signal remains below a specified level  $R$ . The average fade duration is given by:

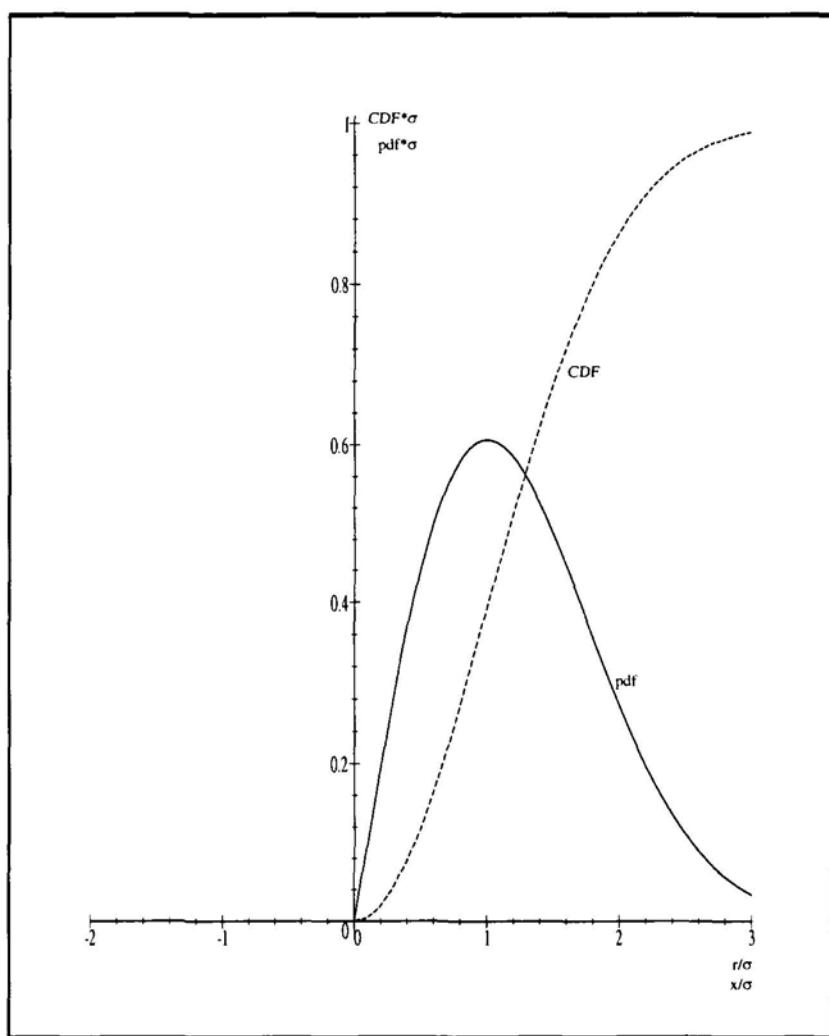
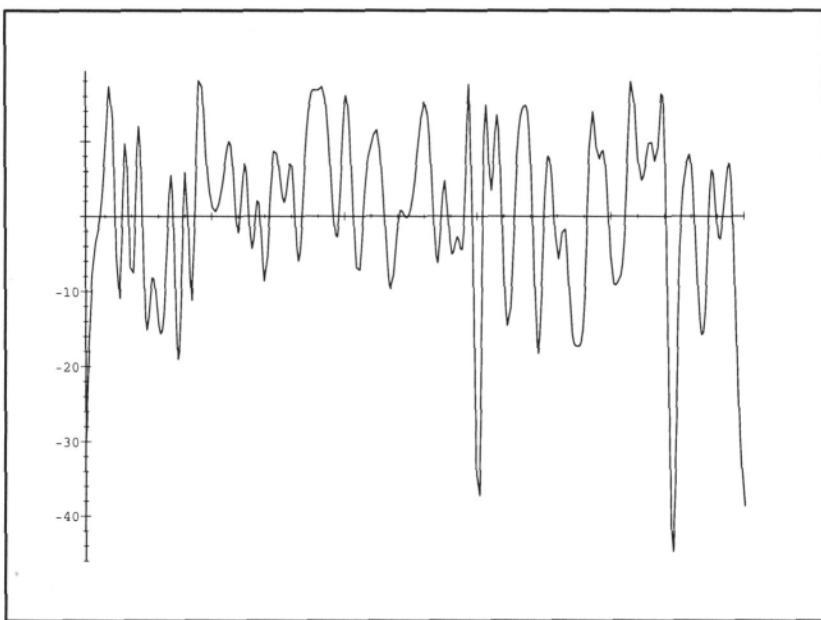


Figure 3.12. Probability distribution function (pdf) and cumulative distribution function (CDF) for a Rayleigh distributed random variable.

$$\tau_R = \frac{e^{\rho^2} - 1}{\rho f_m \sqrt{2\pi}} \quad (3.46)$$

$\tau_R$  determines the average number of bits lost during a single fade.



*Figure 3.13.* Envelope variation of a band-limited Rayleigh distributed random process versus time.

**Example 3.5.** A mobile terminal travels at a speed of 100 Km/hr and receives a signal having a carrier frequency 900 MHz. Calculate:

- (a) The expected level crossing rate (LCR).
- (b) The average fade duration.

Assume that  $\rho = 1$ .

**Solution.** The velocity of the mobile terminal is:

$$v = 100 \times \frac{1000}{3600} = 27.78 \text{ m/sec}$$

The wavelength of the received signal is:

$$\lambda = \frac{c}{f_c} = \frac{3 \times 10^8}{9 \times 10^8} = 0.33 \text{ metres}$$

Using equation 3.32, the maximum Doppler shift is:

$$f_m = \frac{v}{\lambda} = 83.33 \text{ Hz} \quad (3.47)$$

Therefore, using equation 3.45, the expected level crossing rate is:

$$N_R = 76.84 \text{ crossings per second}$$

Using equation 3.46, the average fade duration is:

$$\tau_R = 8.23 \text{ msec}$$

### 3.6.2 Ricean Fading

When there is a dominant non-fading signal component at the receiver, such as a line-of-sight propagation path, the small scale fading distribution becomes Ricean. The resultant wave at the receiver can be represented as:

$$R(t) = A \cos(w_c T) + R_c(t) \cos(w_c T) - R_s(t) \sin(w_c t) \quad (3.48)$$

$A$  is the amplitude of the deterministic non-fading component. While,  $R_c(t)$  and  $R_s(t)$  are Gaussian distributed stationary random processes. The pdf of the Ricean distribution is given by:

$$p(x) = \begin{cases} \frac{x}{\sigma^2} e^{-\frac{x^2 + A^2}{2\sigma^2}} I_0\left(\frac{Ax}{\sigma^2}\right) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.49)$$

Where,

$A$  is a nonnegative value which represents the amplitude of the dominant signal.

$I_o()$  is the modified Bessel function of the first kind and zero order, which is defined as:

$$I_o(x) = \frac{1}{2\pi} \int_0^{2\pi} e^{x \cos \theta} d\theta \quad (3.50)$$

The cumulative distribution function of a Ricean distributed random variable is given by [26]:

$$\text{Prob}(X \leq x) = \begin{cases} 1 - e^{-\frac{A^2 + x^2}{2\sigma^2}} \sum_{m=0}^{\infty} \left(\frac{A}{x}\right)^m I_m\left(\frac{Ax}{\sigma^2}\right) & x \geq 0 \\ 0 & x < 0 \end{cases} \quad (3.51)$$

$I_m$  is the modified Bessel function of the first kind and  $m$ th order.

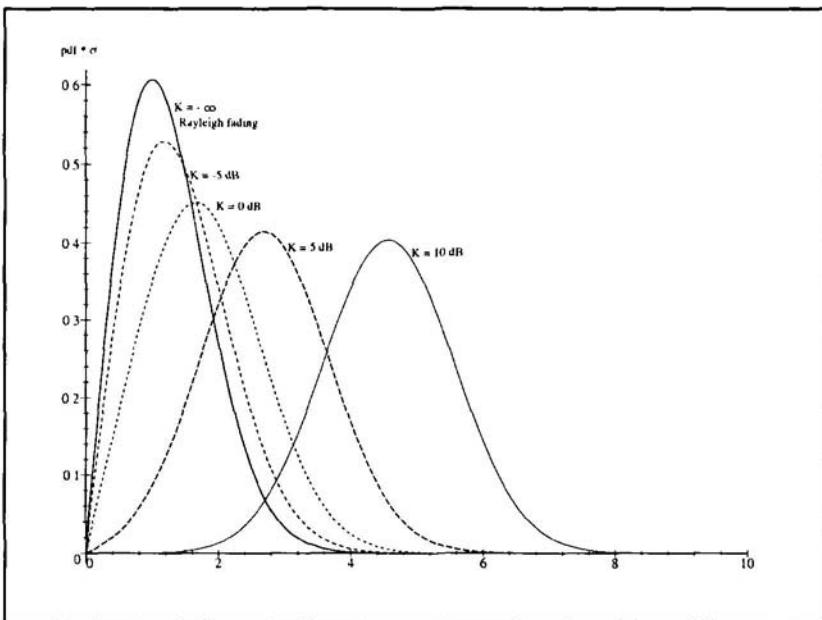


Figure 3.14. Envelope probability density function in a Ricean fading environment.

The Ricean distribution is usually described in terms of a parameter  $K$  [18], known as the Ricean factor, which is the ratio between the dominant signal power and the variance of the multi-path:

$$K(\text{dB}) = 10 \log \left( \frac{A^2}{2\sigma^2} \right) \quad (3.52)$$

The Rayleigh distribution is a special case of the Ricean distribution, as  $A \rightarrow 0$ , the Ricean pdf of equation 3.49 degenerates into the pdf of the Rayleigh distribution given by equation 3.40. On the other hand, as  $A$  becomes larger, the amplitude of the dominant signal increases, the modified Bessel function of the first kind and zero order can be approximated by:

$$I_o(x) \approx \frac{1}{\sqrt{2\pi x}} e^x \quad (3.53)$$

Substituting this value of  $I_o(x)$  into the Ricean pdf of equation 3.49 we get the following pdf distribution, which closely resembles the Gaussian distribution.

$$\begin{aligned}
 p(x) &= \begin{cases} \sqrt{\frac{x}{2\pi\sigma^2 A}} e^{-\frac{(A-x)^2}{2\sigma^2}} & x \geq 0 \\ 0 & x < 0 \end{cases} \\
 &\approx \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(A-x)^2}{2\sigma^2}}
 \end{aligned} \tag{3.54}$$

Figure 3.14 shows the probability density function in a Ricean fading environment for different Ricean factors  $K$ . As  $K \rightarrow -\infty$  ( $A \rightarrow 0$ ), the Ricean distribution degenerates into a Rayleigh distribution. On the other hand, as  $K$  gets larger, the amplitude of the dominant signal becomes larger, the Ricean distribution starts to resemble the Gaussian distribution.

### 3.7 DIVERSITY TECHNIQUES IN WIRELESS COMMUNICATIONS

Diversity techniques are employed in wireless communications systems to improve the channel performance by reducing its sensitivity to noise, interference and fading. Diversity techniques include:

1. Time diversity. This is achieved by interleaving and coding. Interleaving scrambles bursty errors. While, coding adds redundancy to the transmitted signal. Both coding and interleaving were discussed in chapter 2.
2. Frequency diversity. This is achieved by using spread spectrum techniques. Spread spectrum is the topic of chapter 5.
3. Space diversity. This is achieved by using multiple antennas. With multiple antennas, if one antenna is in a fade, it is unlikely that the other antennas will be in a fade as well. This improves the quality of reception.
4. Path diversity. This is achieved by using a rake receiver.

Path diversity is exploited in direct sequence spread spectrum to combat multi-path fading. In direct sequence spread spectrum the data is spread using a pseudorandom sequence having a much higher chip rate. A rake receiver is used to resolve the arriving multi-path components and combine them using a maximal ratio combiner to generate the output signal. Figure 3.15 shows a block diagram of the rake receiver.

Each rake finger is capable of detecting one multi-path component by properly synchronizing the PN code of that rake finger to the delay of one of the arriving multi-path components. The other multi-path components generate Gaussian noise at the output of that rake finger. To be able to resolve multi-path

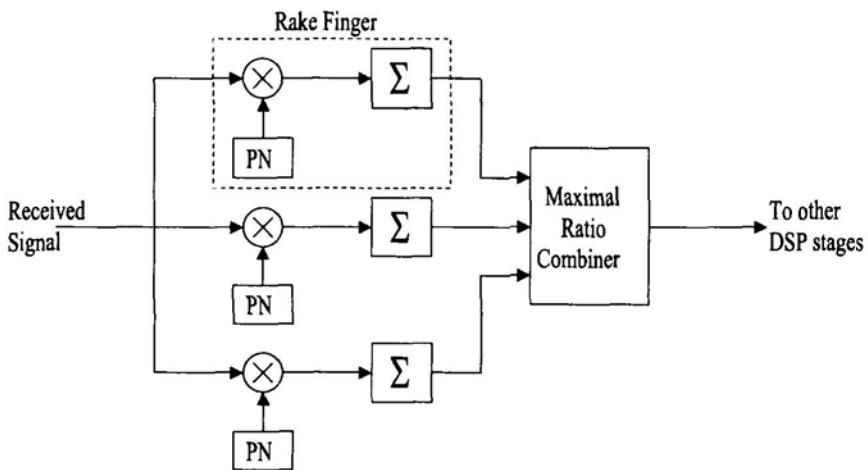


Figure 3.15. A three rake finger receiver.

components the delay between any two adjacent multi-path components needs to be more than the chip duration.

## Chapter 4

# DIGITAL MODULATION SCHEMES

### 4.1 INTRODUCTION

To transmit a signal across a wireless channel it is necessary to modulate this signal onto a carrier. Modulation is defined [27] as the process of imparting the source information onto a band-pass signal with a carrier frequency  $f_c$  by the introduction of amplitude, frequency and/or phase perturbations. The band-pass signal generated is called the modulated signal  $s(t)$ , and the baseband source signal is called the modulating signal  $m(t)$ .

If the modulating baseband signal  $m(t)$  is a digital signal the modulation technique is a digital modulation technique. Digital modulation techniques can be classified into three groups depending on which parameter of the sinusoidal carrier the digital wave modulates. A sinusoidal carrier such as:

$$A_c \cos(w_c T + \phi_c) \quad (4.1)$$

is defined by three parameters, the amplitude  $A_c$ , the frequency  $f_c = w_c/(2\pi)$  and the phase  $\phi_c$ . The subscript “c” denotes the “carrier”.

If the digital signal modulates the amplitude, that modulation scheme is called amplitude modulation. Digital amplitude modulation schemes are presented in section 4.2. If the digital signal modulates the phase, that modulation scheme is called phase modulation. Digital phase modulation schemes (both binary and quadrature) are presented in section 4.3. If the digital signal modulates the frequency, that modulation scheme is called frequency modulation. Digital frequency modulation schemes are presented in section 4.4. Frequency and phase modulation are sometimes grouped together under the name angle modulation. Hybrid modulation schemes also exist, these are a combination of more than one of the previous schemes. Finally, in section 4.5,

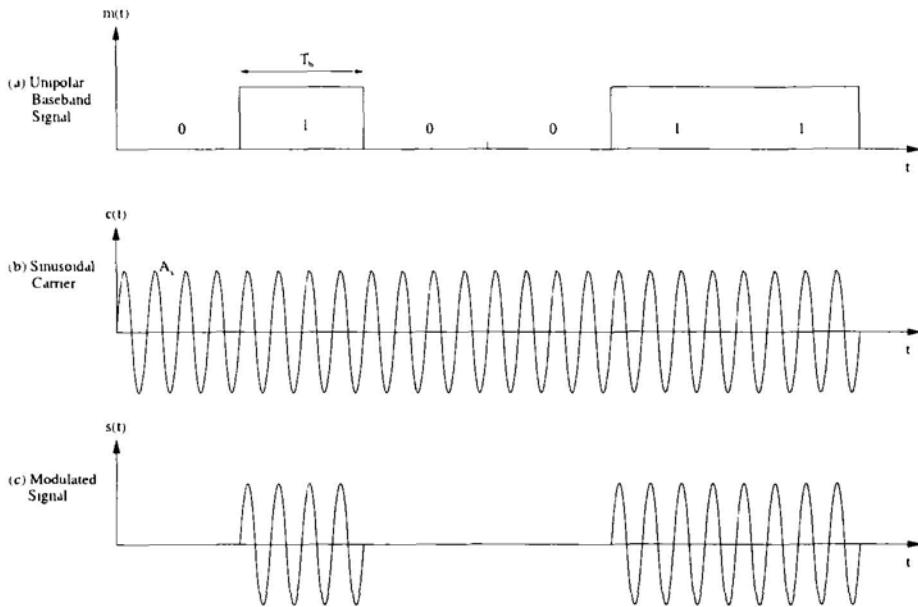


Figure 4.1. Amplitude shift keying (ASK) waveforms.

we compare the power spectral density and the bit error rate performance of different digital modulation schemes.

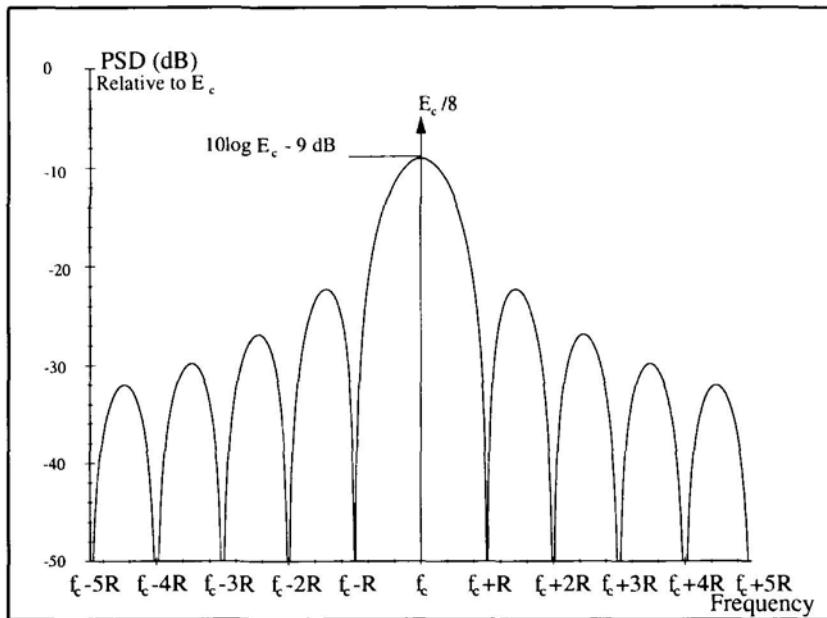
## 4.2 AMPLITUDE SHIFT KEYING

In amplitude shift keying (ASK), which is also known as on-off keying (OOK), the digital signal modulates the amplitude of the sinusoidal carrier. The carrier is switched on and off by a unipolar signal. For a unipolar signal, such as that shown in Figure 4.1.a, logical “1” is represented by a high voltage, while logical “0” is represented by a low (zero) voltage. The modulated signal  $s(t)$  is obtained by multiplying the modulating baseband signal  $m(t)$  by a sinusoidal carrier  $c(t)$ , as shown in Figure 4.1.

The power spectral density of the OOK modulated signal can be derived from that of the unipolar baseband signal and it is found to be equal to [27]:

$$\begin{aligned} G_{OOK}(f) &= \frac{P_c T_b}{8} \left[ \frac{\sin(\pi T_b(f - f_c))}{\pi T_b(f - f_c)} \right]^2 + \frac{P_c}{8} \delta(f - f_c) \\ &\quad + \frac{P_c T_b}{8} \left[ \frac{\sin(\pi T_b(f + f_c))}{\pi T_b(f + f_c)} \right]^2 + \frac{P_c}{8} \delta(f + f_c) \end{aligned} \quad (4.2)$$

$T_b$  is the bit duration,  $f_c$  is the carrier frequency and  $P_c$  is the carrier power, which is related to the carrier amplitude  $A_c$  by the following equation:



**Figure 4.2.** The power spectral density of an amplitude shift keying (ASK) modulated signal such as that given in Figure 4.1.c.  $E_c$  is the bit energy when a logic 1 is transmitted,  $E_c = T_b P_c$ .

$$P_c = \frac{A_c^2}{2} \quad (4.3)$$

Figure 4.2 shows the power spectral density of an OOK modulated signal. The power spectral density is shown for positive frequencies only,  $R = 1/T_b$  is the bit rate. It can be seen from Figure 4.2 that the null-to-null bandwidth is  $2R$ .

The demodulation of an amplitude shift keying modulated signal can be either coherent demodulation, which involves the generation of a sinusoidal signal, which is synchronized in frequency and phase to the received signal, or noncoherent demodulation, which involves the use of an envelope detector. Coherent demodulation is more complex than non-coherent demodulation, but it has superior bit error rate performance. The probability of error (bit error rate) in case of coherent demodulation is given by [27]:

$$P_E = \frac{1}{2} \operatorname{erfc} \left( \sqrt{\frac{E_{bav}}{2N_o}} \right) \quad (4.4)$$

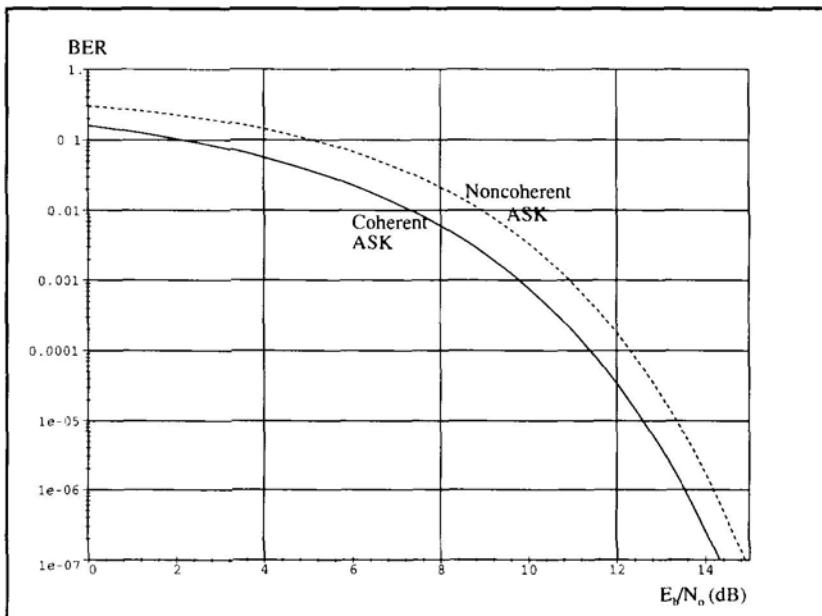


Figure 4.3. Bit error rate for coherent and non-coherent amplitude shift keying (ASK) demodulation, versus  $E_{bav}/N_o$ .

Where,  $E_{bav}$  is the average bit energy.  $N_o$  is the noise power spectral density,  $\text{erfc}(x)$  is the complementary error function, which is defined by equation 2.37. The bit error rate in the case of noncoherent detection is given by[27]:

$$P_E = \frac{1}{2} e^{-E_{bav}/2N_o} \quad (4.5)$$

Equation 4.5 is valid under the following assumptions:

1.  $E_{bav} \gg N_o$ .
2. The bandwidth of the filter proceeding the envelope detector, used for non-coherent demodulation, is equal to the bit rate  $R$ . This is the minimum bandwidth corresponding to a raised cosine filter with  $\alpha = 0$ .

Figure 4.3 shows the bit error rate curves for coherent and noncoherent amplitude shift keying demodulation versus ( $E_{bav}/N_o$ ).

## 4.3 PHASE SHIFT KEYING

In phase shift keying, the digital signal modulates the phase of a sinusoidal carrier. There are many phase shift keying modulation schemes in this section we look at some of these schemes.

### 4.3.1 Binary Phase Shift Keying (BPSK)

In binary phase shift keying (BPSK), the binary digital signal modulates the phase of the sinusoidal carrier. Logic “0” is represented by a carrier having a  $0^\circ$  degree phase shift, hence the transmitted signal is  $A_c \cos(w_c t)$ . While, logic “1” is represented by a carrier having a  $180^\circ$  phase shift, hence the transmitted signal is  $A_c \cos(w_c t + \pi) = -A_c \cos(w_c t)$ .

The modulated signal  $s(t)$  can be generated from the bipolar modulating signal  $m(t)$  of Figure 4.4.a by multiplying the later by the carrier as shown in Figure 4.4. Thus, the BPSK modulated signal  $s(t)$  is given by:

$$s(t) = m(t) A_c \cos(w_c t) \quad (4.6)$$

Where,

$$m(t) = \begin{cases} 1 & \text{for logic 0} \\ -1 & \text{for logic 1} \end{cases} \quad (4.7)$$

Hence,

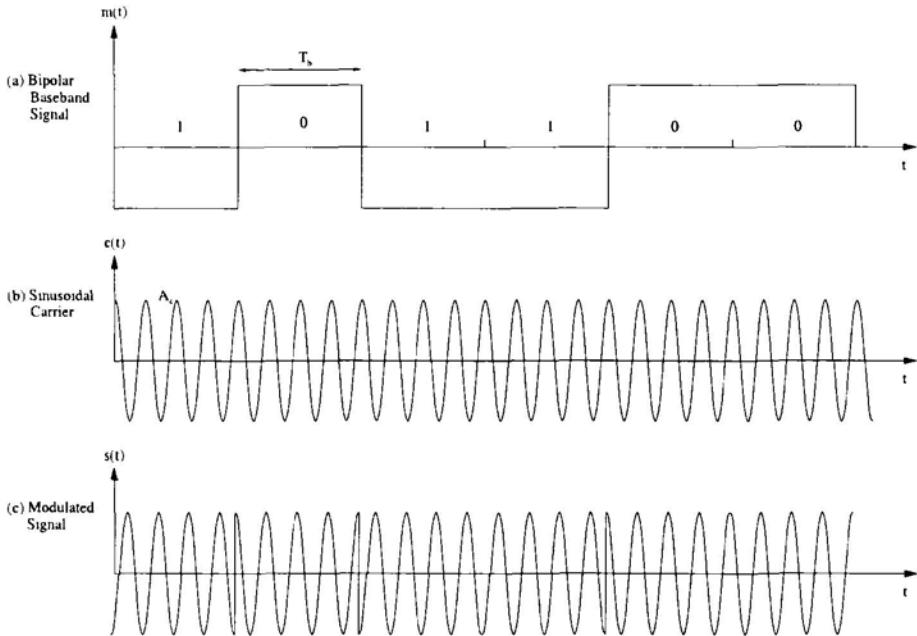
$$s(t) = \begin{cases} A_c \cos(w_c t) & \text{for logic 0} \\ -A_c \cos(w_c t) & \text{for logic 1} \end{cases} \quad (4.8)$$

The power spectral density of the BPSK modulated signal can be derived [27] from that of the bipolar baseband signal and it is found to be given by:

$$\begin{aligned} G_{BPSK}(f) = & \frac{P_c T_b}{2} \left[ \frac{\sin(\pi T_b(f - f_c))}{\pi T_b(f - f_c)} \right]^2 \\ & + \frac{P_c T_b}{2} \left[ \frac{\sin(\pi T_b(f + f_c))}{\pi T_b(f + f_c)} \right]^2 \end{aligned} \quad (4.9)$$

$T_b$  is the bitduration,  $f_c$  is the carrier frequency and  $P_c$  is the carrier power, which is related to the carrier amplitude  $A_c$  by equation 4.3.

Figure 4.23 shows the power spectral density of the BPSK modulated signal. The power spectral density is shown for positive frequencies only. It can be seen from Figure 4.23 that the null-to-null bandwidth for a BPSK modulated signal is  $2R$ . The power spectral density of BPSK is similar to that of OOK with the exception that there is no discrete spectral component at the carrier frequency,  $\pm f_c$ . The presence of a discrete spectral component in the OOK



*Figure 4.4. Binary phase shift keying (BPSK) waveforms.*

modulated signal is related to the fact that the modulating signal for on-off keying is unipolar signal that has a DC component. It is this average DC component that gives rise to the spectral component at the carrier frequency. On the other hand, a bipolar signal, having a zero average DC component, is used as the modulating signal in binary phase shift keying.

Coherent demodulation is used for the demodulation of the BPSK modulated signal. In this case, the bit error rate is given by:

$$P_E = \frac{1}{2} \operatorname{erfc} \left( \sqrt{\frac{E_{bav}}{N_o}} \right) \quad (4.10)$$

Where,  $E_{bav}$  is the average bit energy.  $N_o$  is the noise power spectrum density.  $\operatorname{erfc}(x)$  is the complementary error function, which is defined by equation 2.37. Notice that, for the same bit error rate, BPSK provides a 3 dB improvement over OOK, i.e. for the same bit error rate, OOK requires 3 dB higher signal-to-noise ratio than BPSK. Equation 4.10 is illustrated in Figure 4.24.

### 4.3.2 Differential Binary Phase Shift Keying (DBPSK)

Coherent demodulation must be used in the demodulation of a BPSK signal. This necessitates the use of complex carrier recovery circuits, which provide a reference signal that is synchronized in phase and frequency to the incoming signal to the receiver, and with no phase ambiguity [10]. In a BPSK demodulator, the recovered carrier might not be the required signal  $\cos(w_c t)$ , it might be  $\cos(w_c t + \pi)$  instead. This  $180^\circ$  error, if exists causes a 100% bit error rate.

One way to circumvent the problem of phase ambiguity is to use a differential phase modulation/demodulation scheme, where the bit information is encoded in the relative phase between the signal of the current symbol and that of the previous symbol rather than in the absolute value of the phase. For a differential binary phase shift keying modulation scheme, the phase shift between the phase of the current symbol and that of the previous one is given by the following equation:

$$\Delta\phi = \begin{cases} 0 & \text{logic 0} \\ \pi & \text{logic 1} \end{cases} \quad (4.11)$$

Where,  $\Delta\phi$  is the change in phase of the carrier during the current bit duration relative to that of the previous bit duration. Figure 4.5 shows the block diagram of the differential binary phase shift keying modulator and demodulator. The modulator encodes the incoming bit stream into the phase of the carrier according to the following set of equations:

$$d_k = i_k \oplus d_{k-1} \quad (4.12)$$

$$s(t) = \begin{cases} A \cos(w_c t) & d_k = 0 \\ -A \cos(w_c t) & d_k = 1 \end{cases} \quad (4.13)$$

Where,

$i_k$  is the input bit sequence.

$d_k$  is the differentially encoded bit sequence.

$\oplus$  is the XOR operator.

According to equations 4.12 and 4.13, if  $i_k = 0$ , then  $d_k = d_{k-1}$ , consequently, there is no change in the phase of the carrier. On the other hand, if  $i_k = 1$ , then  $d_k = \bar{d}_{k-1}$ , consequently, the phase of the carrier changes  $180^\circ$ . At the receiver, demodulation is done by multiplying the received signal by a

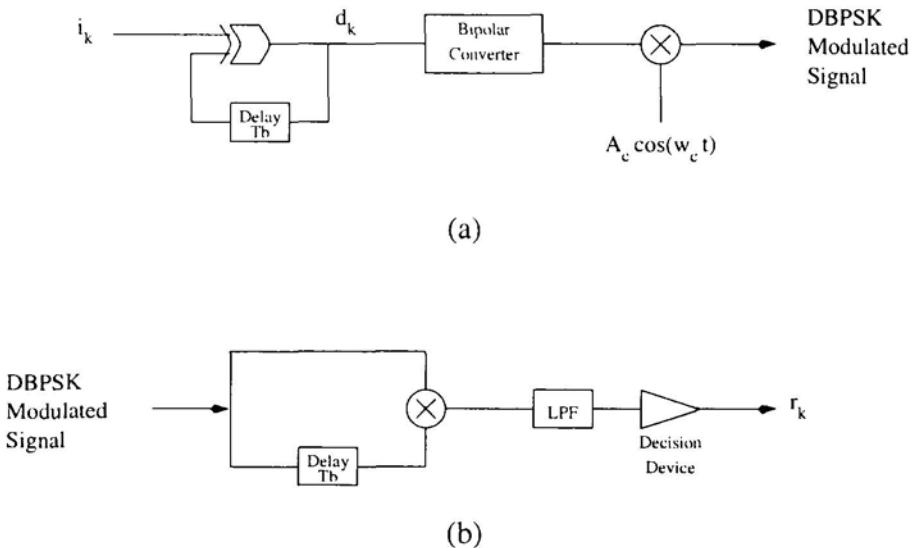


Figure 4.5. Differential binary phase shift keying (DBPSK). (a) Modulator. (b) Demodulator.

one bit delayed version of it. This is known as differential demodulation. The product is low-pass filtered and detected according to the following equation:

$$r_k = \begin{cases} 0 & \text{If } \text{LPF}[s(t)s(t - T_b)] > 0 \\ 1 & \text{If } \text{LPF}[s(t)s(t - T_b)] < 0 \end{cases} \quad (4.14)$$

If there is a  $180^\circ$  phase shift between the received sinusoidal signal of the current bit slot and that of the previous one, then:

$$\text{LPF}[A_c \cos(w_c t) A_c \cos(w_c t + \pi)] = -\frac{A_c^2}{2}.$$

The output of the low-pass filter is negative. Therefore, the received bit  $r_k$  is demodulated as “1”, according to equation 4.14.

On the other hand, if there is no phase shift between the received sinusoidal signal of the current bit slot and that of the previous one, then:

$$\text{LPF}[A_c \cos(w_c t) A_c \cos(w_c t)] = \frac{A_c^2}{2}.$$

The output of the low-pass filter is positive. Therefore, the received bit  $r_k$  is demodulated as “0”, according to equation 4.14.

**Example 4.1.** Assume that the input bit stream to a differential binary phase shift keying (DBPSK) modulator is: 1 0 1 1 0 0 .., and assume that  $d_{-1} = 0$ . Draw the differentially encoded bit stream  $d_k$ , and the DBPSK modulated signal.

**Solution.** The differentially encoded bit stream is determined by equation 4.12, and is shown in Figure 4.6.b. The DBPSK modulated signal is determined by equation 4.13 and is shown in Figure 4.6.c.

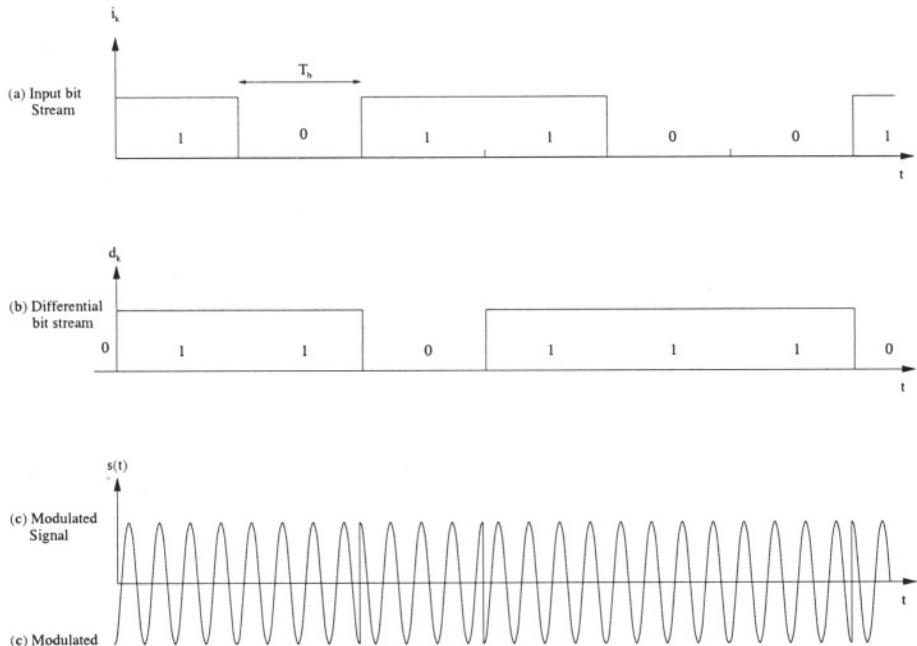


Figure 4.6. Differential binary phase shift keying (DBPSK) waveforms.

The power spectral density of the DBPSK modulated signal is the same as that of the BPSK modulated signal, which is given by equation 4.9, and shown in Figure 4.23.

The bit error rate for the differential demodulation of a DBPSK modulated signal is given by:

$$P_E = \frac{1}{2} e^{-E_{bav}/N_o} \quad (4.15)$$

Where,  $E_{bav}$  is the average bit energy.  $N_o$  is the noise power spectrum density. Notice that, for the same  $E_b/N_o$  the bit error rate performance of DBPSK is worse than that of BPSK. Equation 4.15 is illustrated in Figure 4.24.

It is useful to note that it is also common to use coherent demodulation followed by differential decoding for the detection of a differentially modulated signal. The coherent demodulation gives good BER, while the differential encoding/decoding resolves the phase ambiguity.

### 4.3.3 Quadrature Phase Shift Keying (QPSK)

The modulating signal of a quadrature phase shift keying (QPSK) system is a 4-level digital signal. Each symbol can be represented by two bits (dibits). QPSK modulation involves mapping each dabit into a distinct carrier phase. This mapping process is done according to the space diagram of Figure 4.7 [10]. The modulated signal,  $s(t)$ , for each possible dabit, is given by:

$$s(t) = \begin{cases} \frac{A_c}{\sqrt{2}} \cos(w_c t) + \frac{A_c}{\sqrt{2}} \sin(w_c t) & 00 \\ -\frac{A_c}{\sqrt{2}} \cos(w_c t) + \frac{A_c}{\sqrt{2}} \sin(w_c t) & 10 \\ -\frac{A_c}{\sqrt{2}} \cos(w_c t) - \frac{A_c}{\sqrt{2}} \sin(w_c t) & 11 \\ \frac{A_c}{\sqrt{2}} \cos(w_c t) - \frac{A_c}{\sqrt{2}} \sin(w_c t) & 01 \end{cases} \quad (4.16)$$

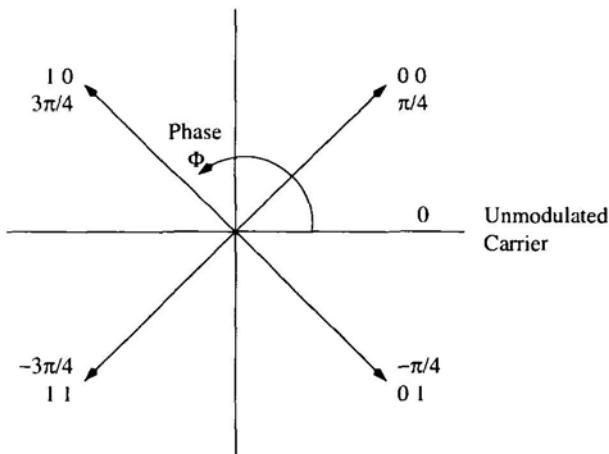


Figure 4.7. Signal-space diagram for quadrature phase shift keying (QPSK) modulation.

According to equation 4.16 and Figure 4.7, the four possible dibits are mapped onto a corresponding phase state in accordance with the Gray code, i.e. every two adjacent phase states only differ by one bit. This property is used to reduce the bit error rate. If an error does occur in one symbol, it is most likely that one of the adjacent phase states is detected by error. In this case, the symbol error (phase state error) will lead to an error in only one bit of the dabit.

Assume each dabit to be represented by the  $IQ$  pair, such that:

$$I \text{ or } Q = \begin{cases} 1 & \text{Logic 0} \\ -1 & \text{Logic 1} \end{cases} \quad (4.17)$$

Therefore, the modulated signal  $s(t)$  of equation 4.16 can be represented by:

$$S_{IQ}(t) = \frac{I}{\sqrt{2}} \cos(w_c t) + \frac{Q}{\sqrt{2}} \sin(w_c t) \quad (4.18)$$

The QPSK modulator shown in Figure 4.8 consists of a demultiplexer that separates the input bit stream into the in-phase stream, which is denoted by I, and the quadrature-phase stream, which is denoted by Q. The I bit stream is delayed by one bit duration ( $T_b$ ), which is equivalent to half a symbol duration ( $T_s$ ), in order to be aligned in time with the Q bit stream. The I and Q streams are multiplied by two sinusoidal signals having a  $90^\circ$  phase shift, as shown in Figure 4.8. The outputs of the two multipliers are then added together to form the QPSK modulated signal. Figure 4.9 shows a QPSK modulated signal for a given input bit stream.

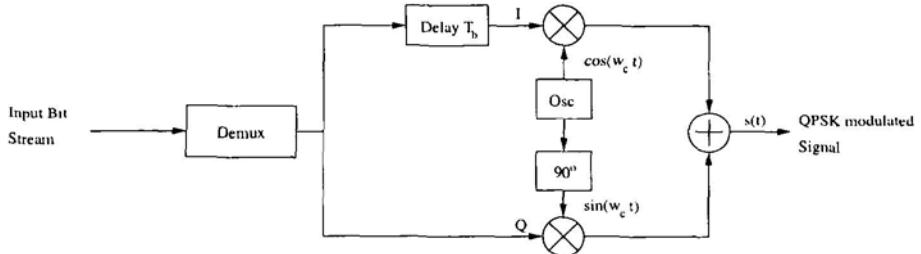


Figure 4.8. Quadrature phase shift keying (QPSK) modulator.

The QPSK demodulator is shown in Figure 4.10. The incoming QPSK modulated signal is multiplied by two sinusoidal signals phase shifted by  $90^\circ$ . The output of the first multiplier is the I stream, the output of the second multiplier is the Q stream. The I and Q streams are multiplexed together to get the output bit stream. It can be seen from Figures 4.8 and 4.10 that the QPSK system can be viewed as two BPSK systems operating in phase quadrature.

The power spectral density of the QPSK modulated signal is given by [27]:

$$\begin{aligned} G_{QPSK}(f) &= P_c T_b \left[ \frac{\sin(2\pi T_b(f - f_c))}{2\pi T_b(f - f_c)} \right]^2 \\ &\quad + P_c T_b \left[ \frac{\sin(2\pi T_b(f + f_c))}{2\pi T_b(f - f_c)} \right]^2 \end{aligned} \quad (4.19)$$

$T_b$  is the bit duration,  $T_b = \frac{1}{2}T_s$ . Where,  $T_s$  is the symbol duration.  $f_c$  is the carrier frequency and  $P_c$  is the carrier power which is related to the carrier amplitude by equation 4.3.

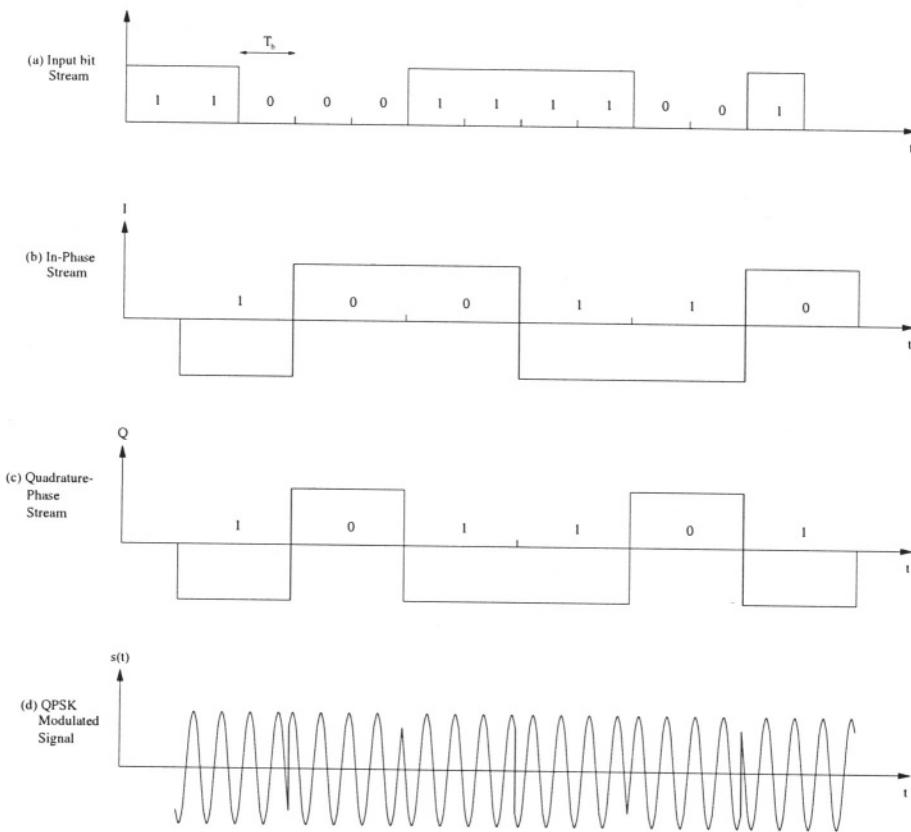


Figure 4.9. Quadrature phase shift keying (QPSK) waveforms.

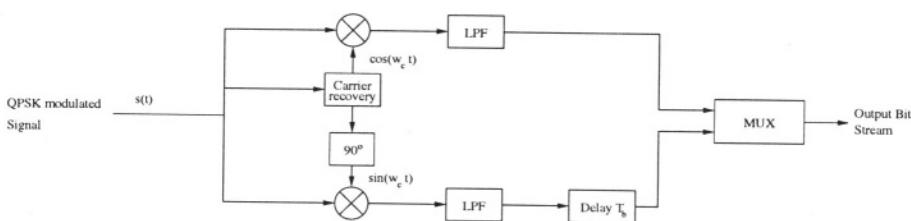


Figure 4.10. Quadrature phase shift keying (QPSK) demodulator.

Figure 4.23 shows the power spectral density of the QPSK modulated signal. The power spectral density is shown for positive frequencies only. It can be seen from Figure 4.23 that the null-to-null bandwidth for a QPSK modulated signal is  $R = 1/T_b$ , which is half that of the BPSK modulated signal. The narrower spectrum means higher spectral efficiency. In fact, a BPSK modulated signal requires a minimum bandwidth of  $R$ , while a QPSK modulated signal requires

a minimum bandwidth of  $R/2$ . The spectral efficiency of a BPSK system is 1 bit/sec/Hz, while that of a QPSK system is 2 bit/sec/Hz. The reason for this is that, in a BPSK system each symbol carries one bit, while in a QPSK system each symbol carries two bits, and Nyquist says we get 1 symbol/second/Hz for a band-pass signal.

Coherent demodulation is used for the demodulation of a QPSK modulated signal. The bit error rate in this case is given by:

$$P_E = \frac{1}{2} \operatorname{erfc} \left( \sqrt{\frac{E_{bav}}{N_o}} \right) \quad (4.20)$$

Where,  $E_{bav}$  is the average bit energy.  $N_o$  is the noise power spectral density,  $\operatorname{erfc}(x)$  is the complementary error function, which is defined by equation 2.37. Equation 4.20 is valid for Gray coded phase states as shown in Figure 4.7. In this case, the bit error rate of a QPSK modulated signal is equal to that of a BPSK modulated signal, which is given by equation 4.10. Equation 4.20 is illustrated in Figure 4.24.

#### 4.3.4 Offset QPSK (OQPSK)

According to Figure 4.9.d, the phase transition between any two adjacent symbols can be  $\{0, \pm 90^\circ, 180^\circ\}$ . These phase transitions lead to envelope fluctuations when the signal is filtered. A  $90^\circ$  phase shift leads to a 70% amplitude fluctuation, while a  $180^\circ$  phase shift leads to a 100% amplitude fluctuation [10]. This amplitude fluctuation leads to an unwanted spectral re-growth and phase modulation as a result of the AM/PM distortion of the nonlinear power amplifiers.

To limit the phase transition between two adjacent symbols to  $90^\circ$  (before filtering), offset quadrature phase shift keying (OQPSK) is used. The I and the Q symbols are not time-aligned, instead they are shifted by  $T_b$ . As a result a transition can never occur simultaneously in both the I and Q streams. Figure 4.11 shows an OQPSK modulated signal and the associated I and Q streams, as well as the OQPSK modulated signal.

The power spectral density and the bit error rate of an OQPSK modulated signal is the same as that of a QPSK modulated signal which is given by equations 4.19 and 4.20 respectively, and shown in Figures 4.23 and 4.24 respectively.

#### 4.3.5 Differential QPSK (DQPSK)

The carrier recovery circuit used in the QPSK demodulator locks to the fourth harmonic of the carrier. This leads to a quarter cycle ambiguity in the recovered carrier [10]. To circumvent such phase ambiguity, the transmitted dibits are

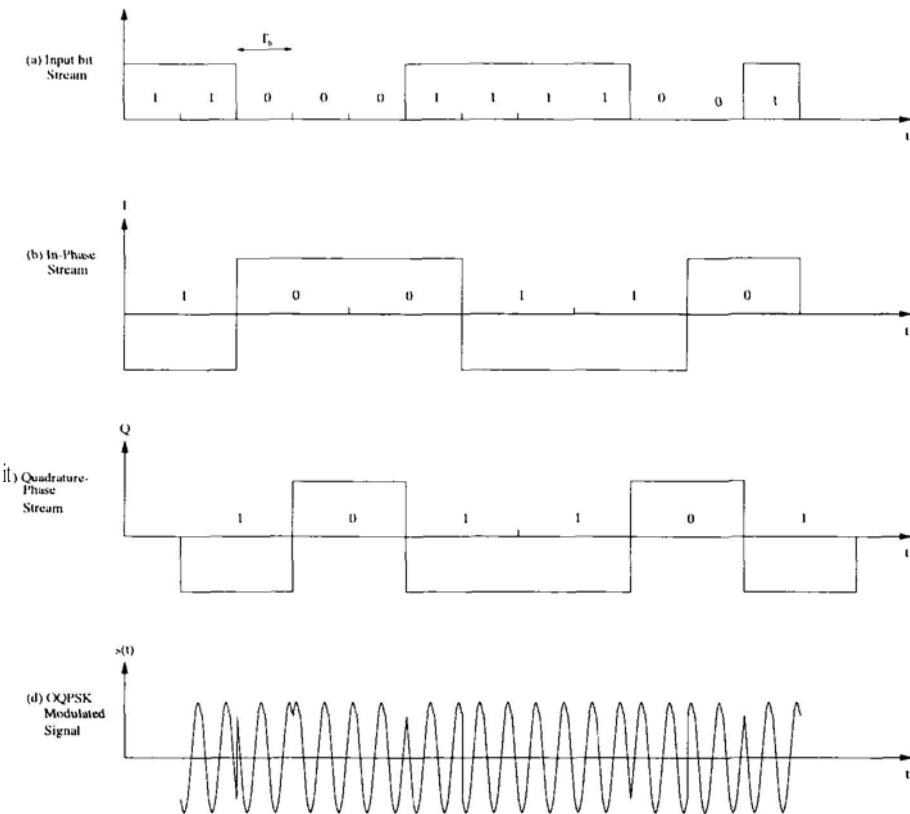


Figure 4.11. Offset quadrature phase shift keying (OQPSK) waveforms.

coded in the phase change between two adjacent phase states, rather than in the absolute value of the phase. This modulation scheme is known as differential quadrature phase shift keying (DQPSK). The phase change depends on the dabit value and is given by the following equation:

$$\Delta\phi = \begin{cases} 0^\circ & \text{If dabit is 00} \\ 90^\circ & \text{If dabit is 10} \\ 180^\circ & \text{If dabit is 11} \\ -90^\circ & \text{If dabit is 01} \end{cases} \quad (4.21)$$

The incoming  $I_n$  and  $Q_n$  bit streams are encoded into the  $D_n$  and  $E_n$  bit streams according to the following equations [10]:

$$D_n = (\overline{I_n \oplus Q_n})(I_n \oplus D_{n-1}) + (I_n \oplus Q_n)(Q_n \oplus E_{n-1}) \quad (4.22)$$

$$E_n = (\overline{I_n \oplus Q_n})(I_n \oplus E_{n-1}) + (I_n \oplus Q_n)(I_n \oplus D_{n-1}) \quad (4.23)$$

The  $+$  operator in equations 4.22 and 4.23 denotes the logic OR operator, while the  $\oplus$  operator denotes the logic XOR operator. Table 4.1 gives the relationship between the differentially encoded dabit of time slot  $n$  ( $D_n E_n$ ) and that of time slot  $(n - 1)$  ( $D_{n-1} E_{n-1}$ ) as a function of the input dubits ( $I_n Q_n$ ).

*Table 4.1.* Differential encoding for quadrature phase shift keying.

$I_n$	$Q_n$	$D_n$	$E_n$
0	0	$D_{n-1}$	$E_{n-1}$
0	1	$\bar{E}_{n-1}$	$D_{n-1}$
1	0	$E_{n-1}$	$\bar{D}_{n-1}$
1	1	$\bar{D}_{n-1}$	$\bar{E}_{n-1}$

During time slot  $n$ , the transmitted signal can be represented by:

$$D_n \cos(w_c t) + E_n \sin(w_c t) = \sqrt{2} \cos(w_c t - \phi_n) \quad (4.24)$$

Where,  $D_n$  and  $E_n$  are represented by +1 and -1 for the 0 and 1 logical states respectively. Furthermore,  $\phi_n$  is given by:

$$\sqrt{2} e^{j\phi_n} = D_n + j E_n \quad (4.25)$$

The differential phase transition between two consecutive symbols is given by:

$$\Delta\phi_n = \phi_{n-1} - \phi_n \quad (4.26)$$

At the receiver side, the DQPSK signal is demodulated as shown in Figure 4.12, which can be viewed as two DBPSK demodulators operating in parallel, with one of them introducing a phase shift of  $135^\circ$  in the delayed signal, while the other introduces a phase shift of  $-135^\circ$  in the delayed signal.

The output of the multiplier of the top branch of Figure 4.12 is given by:

$$\begin{aligned} & 2 \cos(w_c t - \phi_n) \cos(w_c t - \phi_{n-1} + 135^\circ) \\ &= \cos(2w_c t - (\phi_n + \phi_{n-1}) + 135^\circ) + \cos(\Delta\phi_n - 135^\circ) \end{aligned} \quad (4.27)$$

Hence, the output of the LPF filter of the top branch of Figure 4.12 is given by:

$$\cos(\Delta\phi_n - 135^\circ) \quad (4.28)$$

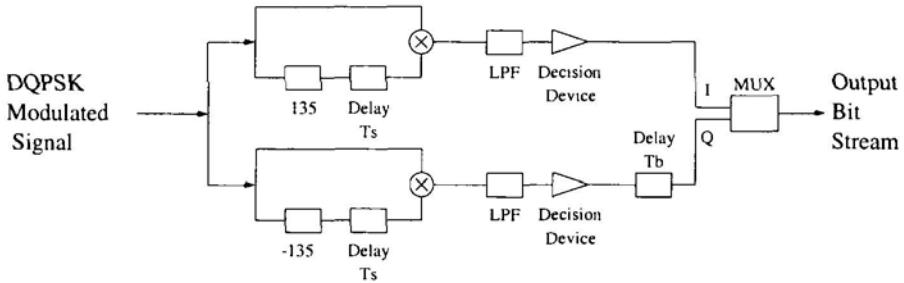


Figure 4.12. Differential quadrature phase shift keying (DQPSK) demodulator.

If  $I_n$  is logical 1, this quantity is positive, on the other hand if  $I_n$  is logical 0, this quantity is negative. Similarly, the output of the LPF of the bottom branch of Figure 4.12 is given by:

$$\cos(\Delta\phi_n + 135^\circ) \quad (4.29)$$

If  $Q_n$  is logical 1, this quantity is positive, on the other hand if  $Q_n$  is logical 0, this quantity is negative.

**Example 4.2.** Assume that the  $I_n Q_n$  dabit during symbol  $n$  is 10. Find the phase change in the DQPSK signal according to equations 4.22 and 4.23 and Table 4.1. Show that the demodulator of Figure 4.12 correctly demodulates the DQPSK modulated signal.

**Solution.** According to equations 4.22 and 4.23, the differentially encoded signal during symbol  $n$  is related to that during symbol  $(n-1)$  by the following equations:

$$D_n = E_{n-1}, \quad E_n = \overline{D_{n-1}}$$

If the DQPSK signal during symbol  $(n-1)$  is given by:

$$\text{DQPSK}_{n-1}(t) = D_{n-1} \cos(w_c t) + E_{n-1} \sin(w_c t) \quad (4.30)$$

Therefore, the DQPSK signal during symbol  $n$  is given by:

$$\text{DQPSK}_n(t) = E_{n-1} \cos(w_c t) - D_{n-1} \sin(w_c t) \quad (4.31)$$

Notice that, the phase transition between the signal during symbol  $(n-1)$  and that during symbol  $n$  is  $90^\circ$  as specified by equation 4.21 for a 10 dabit. At the receiver, the output of the  $I$  differential demodulator, during symbol  $n$ , is given by:

$$\begin{aligned} I &= \text{LPF}[\Phi_{135}(\text{DQPSK}_{n-1}(t))\text{DQPSK}_n(t)] \\ &= \frac{1}{2\sqrt{2}} [D_n^2 + E_n^2] = \frac{\sqrt{2}}{2} \longrightarrow \text{+ve} \longrightarrow I_n \text{ is logic 1} \quad (4.32) \end{aligned}$$

$\Phi_{135}$  denotes a wideband  $135^\circ$  phase shifter. In arriving at the result of equation 4.32, we made use of the following facts:

- The LPF filters out all the frequency components at  $2f_c$ .
- Since,  $D_n, E_n \in \{1, -1\}$ . Therefore,  $D_n^2 = E_n^2 = 1$ .

Similarly, the output of the  $Q$  differential demodulator, during symbol  $n$ , is given by:

$$\begin{aligned} Q &= \text{LPF}[\Phi_{-135}(\text{DQPSK}_{n-1}(t))\text{DQPSK}_n(t)] \\ &= \frac{-1}{2\sqrt{2}} [D_n^2 + E_n^2] = -\frac{\sqrt{2}}{2} \longrightarrow \text{-ve} \longrightarrow Q_n \text{ is logic 0} \quad (4.33) \end{aligned}$$

The power spectral density of a DQPSK modulated signal is the same as that of a QPSK modulated signal which is given by equation 4.19 and illustrated in Figure 4.23.

The bit error rate of a DQPSK modulated signal is given approximately by [10]:

$$P_E \approx e^{-(E_{bav}/N_o)(2-\sqrt{2})} \quad (4.34)$$

Where,  $E_{bav}$  is the average bit energy, and  $N_o$  is the noise power spectral density. Equation 4.34 assumes that the bandwidth of the band-pass filters used in the DQPSK system is half the bit rate. In this case,  $E_{bav}/N_o$  is related to the carrier-to-noise ratio ( $C/N$ ) by the following relation:

$$E_{bav}/N_o = \frac{1}{2} C/N \quad (4.35)$$

Equation 4.34 is illustrated in Figure 4.24. Notice that, the bit error rate performance of DQPSK is worse than that of QPSK and DBPSK.

### 4.3.6 $\pi/4$ - DQPSK

Unlike conventional DQPSK that has four possible phase states,  $\pi/4$ -DQPSK has eight possible phase states.  $\pi/4$ -DQPSK is a differential phase shift keying modulation scheme, the transmitted bits are grouped into two-bit (dibits) symbols that are then encoded into the relative phase change from one symbol

to the next. The phase change between two adjacent symbols is related to the modulating dabit, using a Gray code representation, according to the following equation [28]:

$$\Delta\phi = \begin{cases} +45^\circ & \text{If dabit is 00} \\ +135^\circ & \text{If dabit is 01} \\ -135^\circ & \text{If dabit is 11} \\ -45^\circ & \text{If dabit is 10} \end{cases} \quad (4.36)$$

$\pi/4$ -DQPSK was selected as the digital modulation scheme used in the North American digital cellular standard (IS-136).

One advantage of the  $\pi/4$ -DQPSK modulation scheme is that it exhibits symmetric phase transitions [29]. The phase shift between one dabit symbol to the next is  $\pm\pi/4$  or  $\pm3\pi/4$ . Thus, it avoids the  $\pi$  phase shift of the DQPSK modulation scheme, which leads to a lower envelope fluctuations in the filtered modulated signal. Hence, the  $\pi/4$ -DQPSK modulation scheme suffers less spectral re-growth, due to the nonlinear power amplifier preceding the antenna, than the DQPSK modulation scheme.

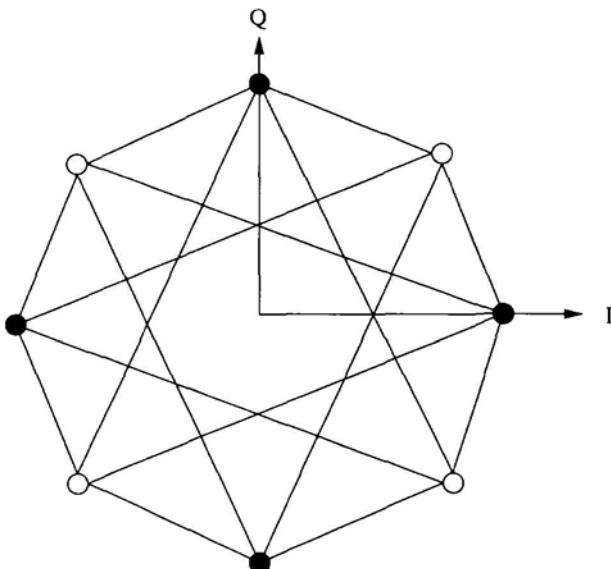


Figure 4.13. The space diagram of a  $\pi/4$ -DQPSK system showing all allowable phase state transitions.

Figure 4.13 shows the space diagram of a  $\pi/4$ -DQPSK system. The eight allowable phase states are divided into two groups of four states each. During a certain symbol, the phase belongs to one of the two groups. During the next symbol, the phase belongs to the other group, and so on. The lines joining two

phase-states, in Figure 4.13, show an allowed phase state transition between two adjacent dabit symbols.

The in-phase  $I_n$ , and quadrature-phase  $Q_n$  components for a  $\pi/4$ -DQPSK modulated signal, during symbol  $n$ , are related to those during symbol  $n - 1$  by the following equations:

$$I_n = I_{n-1} \cos(\Delta\phi) - Q_{n-1} \sin(\Delta\phi) \quad (4.37)$$

$$Q_n = I_{n-1} \sin(\Delta\phi) + Q_{n-1} \cos(\Delta\phi) \quad (4.38)$$

Where,  $\Delta\phi$  is given by equation 4.36. The  $\pi/4$ -DQPSK modulated signal can be detected using a coherent detector, a differential detector [30], or a limiter-discriminator detector [31]. The coherent detector is of higher complexity than the differential and limiter-discriminator detectors because of the carrier recovery circuitry and the linear receiver requirement. Furthermore, in a fading environment, coherent detection results in a higher BER [31], [32]. Figure 4.14 [31] shows the block diagram of the carrier recovery circuitry used to extract the carrier of a  $\pi/4$ -DQPSK modulated signal. The received signal is raised to the fourth power and then divided by four, in the frequency domain, to get the recovered carrier. This gives rise to a  $90^\circ$  phase ambiguity in the recovered carrier. However, because the information is stored in the phase difference and not in the absolute value of the phase, the signal can be correctly demodulated. The multiplication by a clock having half the symbol rate is necessary because the four phase states are used every other symbol.

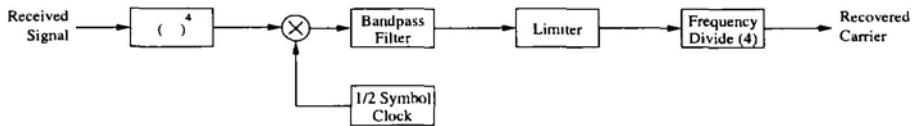


Figure 4.14. The carrier recovery circuit for a  $\pi/4$ -DQPSK modulated signal. Source [31].

Figure 4.15 [31], shows a block diagram of the limiter-discriminator detector. This is the easiest method for detection of a  $\pi/4$ -DQPSK modulated signal since it uses the familiar FM receiver technology. Hence, it has the advantage of being used to detect both analog FM modulated signals, used in the North American AMPS standard, and  $\pi/4$ -DQPSK modulated signals, used in IS-136. Figure 4.16 [30], shows a block diagram of a differential  $\pi/4$ -DQPSK detector. The differential demodulator requires a linear receiver [31] to properly detect the  $\pi/4$ -DQPSK signal, this makes it of higher complexity than the limiter-discriminator detector, but of lower complexity than the coherent detector.

$\pi/4$ -DQPSK has the same frequency spectrum as QPSK, which is given by equation 4.19, and illustrated in Figure 4.23. The bit error rate depends

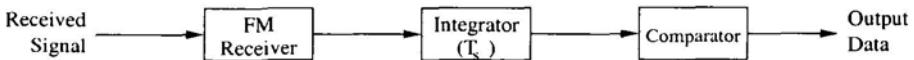


Figure 4.15. Limiter discriminator detector for a  $\pi/4$ -DQPSK modulated signal. Source [31].

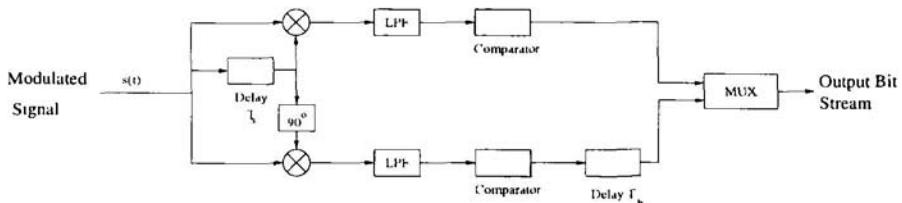


Figure 4.16. Differential detector for a  $\pi/4$ -DQPSK modulated signal. Source [30].

on the type of detector used. Figure 4.17 shows the bit error rate for the three detectors presented earlier in this subsection, in a static environment (no fading) and using a square-root raised cosine filter with a roll-off factor  $\alpha = 0.35$  in both the transmitter and receiver. Also shown in Figure 4.17 is the bit error rate performance of a coherently detected QPSK modulation scheme, which is given by equation 4.20. Notice that, the bit error rate performance of the limiter-discriminator detector and that of the differential detector are almost the same. The bit error rate performance of the coherent detector is better than that of the limiter-discriminator and differential detectors, however, it is lower than that of a coherently detected QPSK modulated signal. In a fading environment, as mentioned earlier, the bit error rate performance of the coherent detector is worse than that of the limiter-discriminator and differential detectors.

## 4.4 FREQUENCY SHIFT KEYING

### 4.4.1 Minimum Shift Keying (MSK)

Digital phase modulation schemes such as those presented in the last section undergo abrupt phase changes at the symbol boundary, which in turn leads to discontinuity in the unfiltered modulated signal. This has two disadvantages. First, an abrupt change in the value of the input signal gives rise to high frequency components, which causes adjacent channel interference (ACI). Second, these high frequency components can be suppressed by filtering the modulated signal. However, this filtering process leads to amplitude (envelop) fluctuations so that when the filtered signal passes through the power amplifier, which is a highly non-linear device, the amplitude fluctuations in the input to the amplifier lead to spectral re-growth at the output from the amplifier. This leads to adjacent channel interference.

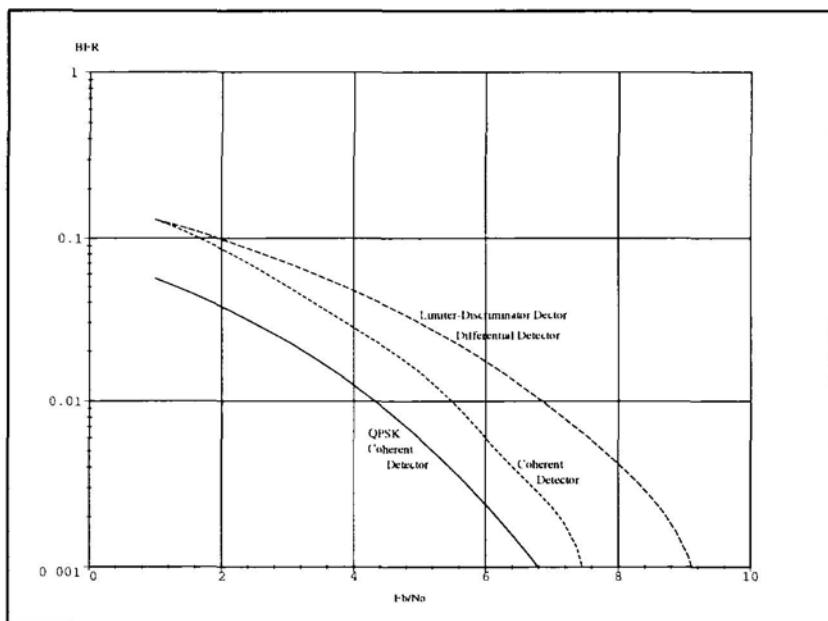


Figure 4.17. The bit error rate of a  $\pi/4$ -DQPSK modulation scheme. Source [31].

To avoid these problems a modulation scheme is needed such that the phase changes continuously from one symbol to the next. One such modulation scheme is Minimum Shift Keying (MSK). MSK is a frequency modulation scheme where the frequency of the modulated signal changes from one symbol to the next according to the input bit stream.

The two frequencies that can be transmitted by an MSK modulator are related to the carrier frequency  $f_c$  and the bit duration  $T_b$  by the following equations:

$$f_1 = f_c + \frac{1}{4T_b} \quad (4.39)$$

$$f_2 = f_c - \frac{1}{4T_b} \quad (4.40)$$

The separation between  $f_1$  and  $f_2$  is the minimum separation necessary to ensure orthogonality of the two waveforms. This explains the name Minimum Shift Keying.

During each bit interval the phase of the carrier advances or retards by  $90^\circ$ , depending on whether frequency  $f_1$  or  $f_2$  was transmitted. Furthermore,

unlike phase shift keying modulation schemes the phase of the transmitted signal never changes abruptly, rather it changes continuously during the bit interval. Figure 4.18.b shows how the phase of the carrier changes for the input bit stream given in Figure 4.18.a. Notice that, for logic 0 frequency  $f_1$  is transmitted, while for logic 1 frequency  $f_2$  is transmitted.

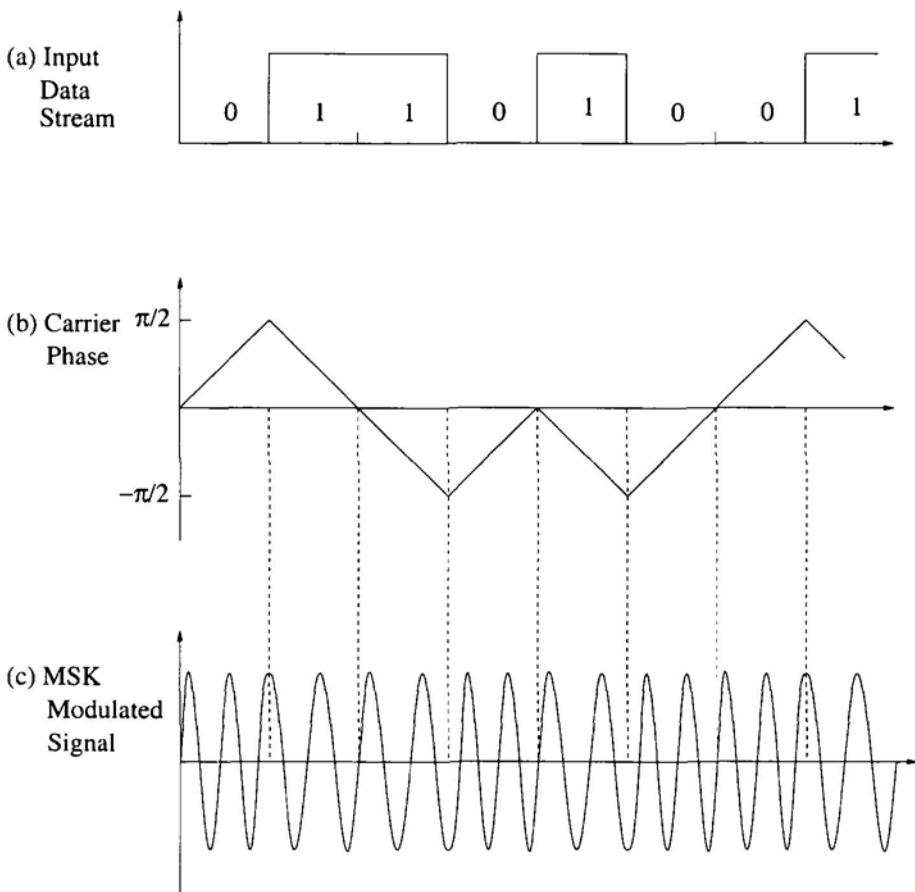


Figure 4.18. Minimum Shift Keying (MSK) waveforms.

The simplest way to implement an MSK modulator is to use a voltage controlled oscillator (VCO) as shown in Figure 4.19.a. The input bit stream controls the output frequency of the VCO. If the input signal is logic 0, the output signal of the VCO has a frequency  $f_1$ . On the other hand, if the input signal is logic 1, the output signal of the VCO has a frequency  $f_2$ .

During bit interval  $n$  ( $nT_b < t < (n+1)T_b$ ), the phase of the transmitted signal is given by:

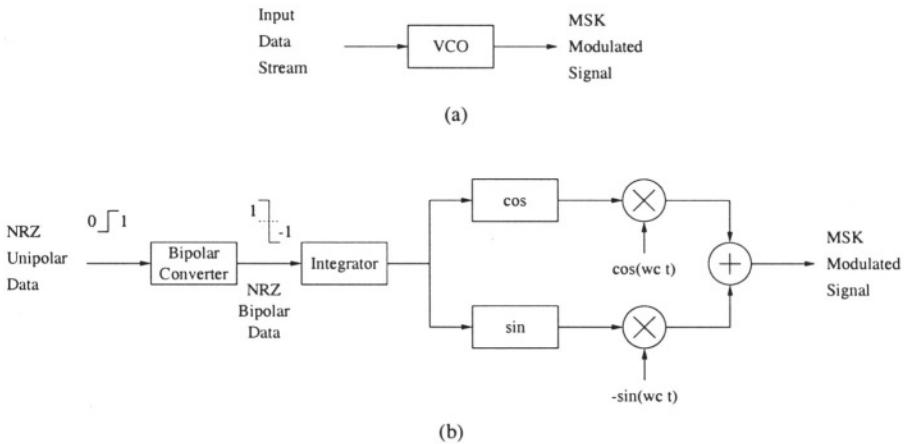


Figure 4.19. Minimum Shift Keying (MSK) Modulator.

$$\phi(t) = \phi_n + \frac{\pi}{2} b_n \frac{t - nT_b}{T_b} \quad (4.41)$$

Where,  $\phi_n$  is the phase of the transmitted signal at  $t = nT_b$ .  $\phi_n$  depends on the previously transmitted data according to the following equation:

$$\phi_n = \sum_{i=0}^{n-1} \frac{\pi}{2} b_i \quad (4.42)$$

$b_i$  depends on the logic value of the bit transmitted during the  $i$ th interval, according to the following equation:

$$b_i = \begin{cases} 1 & \text{Logic 0} \\ -1 & \text{Logic 1} \end{cases} \quad (4.43)$$

The MSK modulated signal can be expressed as:

$$\begin{aligned} s(t) &= \cos [2\pi f_c t + \phi(t)] \\ &= \cos 2\pi f_c t \cos \phi(t) - \sin 2\pi f_c t \sin \phi(t) \end{aligned} \quad (4.44)$$

Where,  $\phi(t)$  is the integration of the NRZ bipolar data stream ( $b_i$ ). Figure 4.19.b shows an MSK modulator implemented according to equation 4.44.

Figure 4.20 shows a block diagram of one possible implementation of an MSK demodulator. A carrier recovery circuit is used to recover the carrier ( $\cos w_c t$ ), and generate the quadrature signal of the carrier ( $\sin w_c t$ ). The received signal is multiplied by both the recovered carrier and its quadrature signal. The product of each multiplier is filtered by a low-pass filter to eliminate

the frequency components at double the carrier frequency. The output of the upper low-pass filter is the in-phase component ( $I$ ) of the received signal, while the output of the lower low-pass filter is the quadrature-phase component ( $Q$ ). Using  $I$  and  $Q$ , it is possible to calculate the instantaneous phase of the received signal. Differentiating this phase with respect to time gives the instantaneous frequency, and hence the digital bit stream transmitted across the channel.

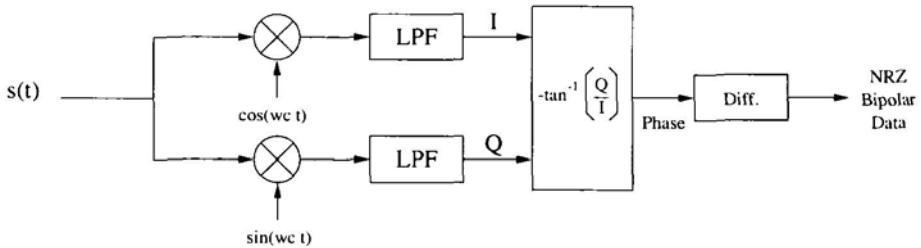


Figure 4.20. Minimum Shift Keying (MSK) Demodulator.

The power spectral density of the MSK modulated signal is given by [10]:

$$G_{\text{MSK}} = 4P_c T_b \left\{ \frac{[1 + \cos 4\pi(f - f_c)T_b]}{\pi^2 [1 - 16T_b^2(f - f_c)^2]^2} + \frac{[1 + \cos 4\pi(f + f_c)T_b]}{\pi^2 [1 - 16T_b^2(f + f_c)^2]^2} \right\} \quad (4.45)$$

Figure 4.23 shows the power spectral density of the MSK modulated signal. The power spectral density is shown for positive frequencies only. It can be seen from Figure 4.23 that the null-to-null bandwidth for an MSK modulated signal is  $R = 1.5/T_b$ . This is one-and-a-half times larger than that of a QPSK modulated signal. Despite the wider main lobe of an MSK modulated signal, yet the spectrum falls at a rate of  $f^{-4}$  (40 dB per decade), which is double the rate at which the spectrum of a QPSK modulated signal falls by. The spectrum of a QPSK modulated signal falls at a rate of  $f^{-2}$  (20 dB per decade). The spectral efficiency of an MSK modulated signal is 2 b/sec/Hz.

The bit error rate of an MSK modulated signal is same as that of a QPSK modulated signal, which is given by:

$$P_E = \frac{1}{2} \operatorname{erfc} \left( \sqrt{\frac{E_{bav}}{N_o}} \right) \quad (4.46)$$

Where,  $E_{bav}$  is the average bit energy, and  $N_o$  is the noise power spectral density,  $\operatorname{erfc}(x)$  is the complementary error function, which is defined by equation 2.37. Equation 4.46 is illustrated in Figure 4.24.

#### 4.4.2 Gaussian Minimum Shift Keying (GMSK)

Gaussian Minimum Shift Keying (GMSK) is a digital frequency modulation scheme similar to MSK, except that the modulating signal (the bipolar NRZ data stream) is filtered using a Gaussian filter before being applied to the integrator. Since both the Gaussian filter and the integrator are linear time invariant operators, their order can be reversed without affecting the functionality of the system. A low-pass Gaussian filter has a frequency response given by:

$$H_{\text{Gaus}}(f) = K e^{\frac{\ln 2}{2} \frac{f^2}{B^2}} \quad (4.47)$$

Where,  $B$  is the 3-dB bandwidth.  $H_{\text{Gaus}}(f)$  has an impulse response given by:

$$h_{\text{Gaus}}(t) = K \sqrt{\frac{2\pi}{\ln 2}} B e^{-\frac{2\pi^2 B^2 t^2}{\ln 2}} \quad (4.48)$$

Let a rectangular pulse with width  $T_b$  be applied to a Gaussian filter having a 3-dB bandwidth  $B$ . Knowing that:

$$\int_0^x e^{-u^2} du = \frac{\sqrt{\pi}}{2} \operatorname{erf}(x) \quad (4.49)$$

and that  $\operatorname{erf}(x)$  is an odd function, i.e.  $\operatorname{erf}(-x) = -\operatorname{erf}(x)$ , it is possible to show that the pulse response of a Gaussian low-pass filter, for a pulse extending between  $T_b/2$  to  $-T_b/2$ , is given by:

$$P_{\text{Gaus}}(t) = \frac{K}{2} \left\{ \operatorname{erf} \left[ \pi \sqrt{\frac{2}{\ln 2}} \left( t + \frac{T_b}{2} \right) B \right] - \operatorname{erf} \left[ \pi \sqrt{\frac{2}{\ln 2}} \left( t - \frac{T_b}{2} \right) B \right] \right\} \quad (4.50)$$

Figure 4.21 shows the pulse response of the Gaussian filter for different values of  $BT_b$ . GMSK improves the spectrum of the MSK modulation scheme by making the spectrum more compact.

A GMSK modulator is similar to an MSK modulator except that it has a Gaussian filter inserted before (or after) the phase integrator. Figure 4.22 shows a block diagram of a GMSK modulator. The GMSK demodulator is identical to the MSK demodulator which is shown in Figure 4.20.

### 4.5 DIGITAL MODULATION TECHNIQUES COMPARISONS

Figure 4.23 shows the power spectrum density of different frequency and phase digital modulation schemes. Notice that, binary phase shift keying modulation schemes such BPSK and DBPSK have a null-to-null bandwidth of twice the

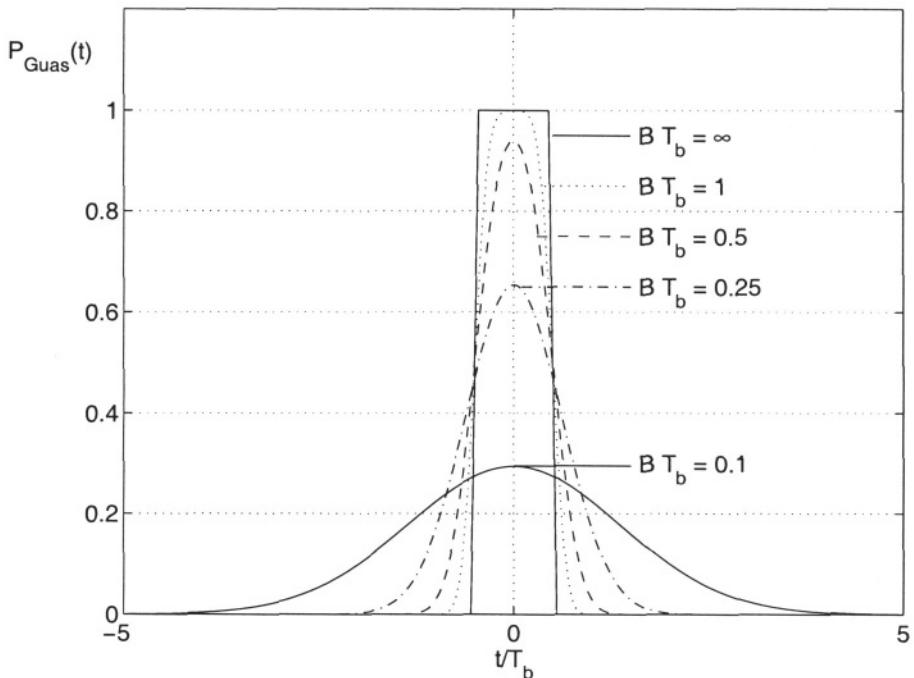


Figure 4.21. Pulse response of a Gaussian low-pass filter.

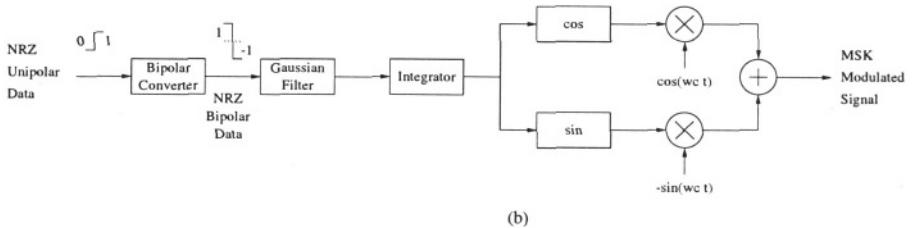


Figure 4.22. Gaussian Minimum Shift Keying (GMSK) Modulator.

bit rate, and decay with frequency at a rate of 20 dB/decade. Quadrature phase shift keying modulation schemes such as QPSK, DQPSK and OQPSK have a null-to-null bandwidth equal to the bit rate, and decay with frequency at a rate of 20 dB/decade. MSK has a null-to-null bandwidth of one-and-a-half times the bit rate, but decays with frequency at a rate of 40 dB/decade.

Figure 4.24 shows the bit error rate performance for different frequency and phase digital modulation schemes. The coherent modulation schemes such as BPSK, QPSK, OQPSK and MSK have the best bit error rate performance.

Differential modulation schemes have lower bit error rate performance, with DQPSK worse than DBPSK.

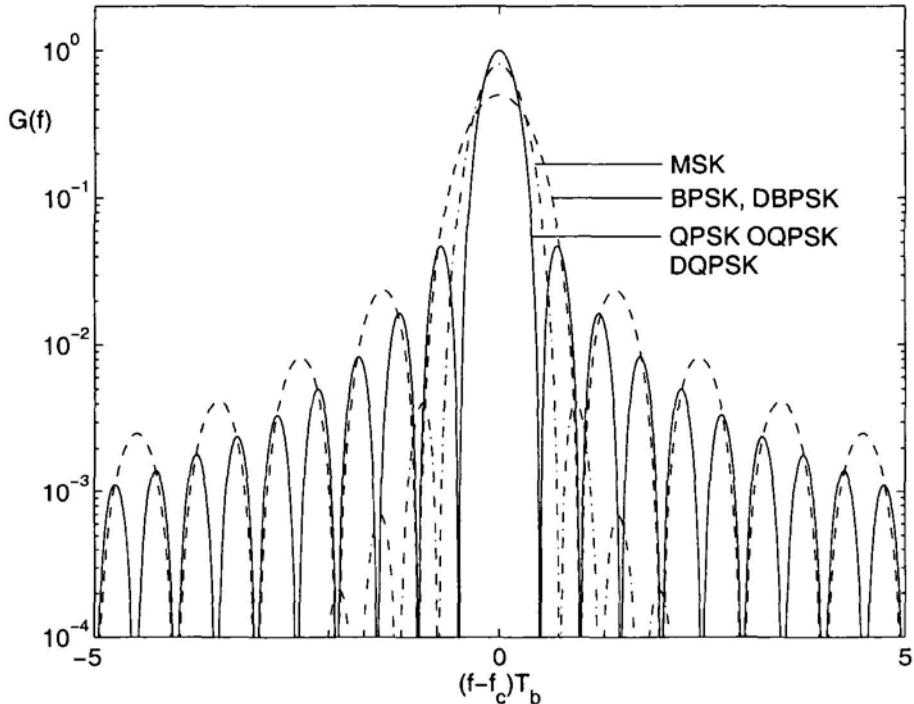


Figure 4.23. Power spectrum density for digital modulation schemes.

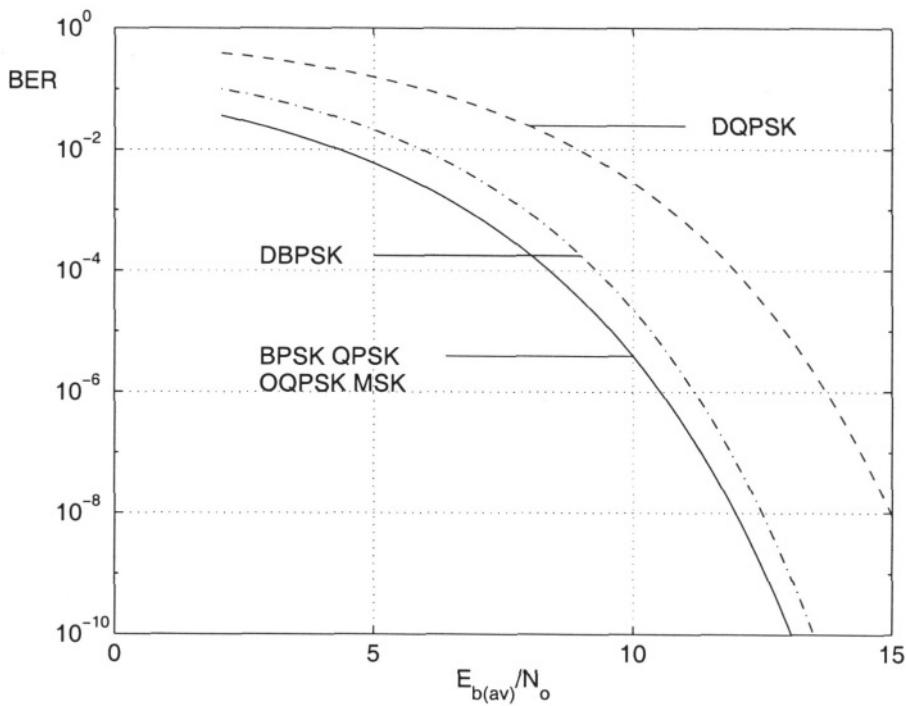


Figure 4.24. Bit error rate for digital modulation schemes.

# Chapter 5

## SPREAD SPECTRUM

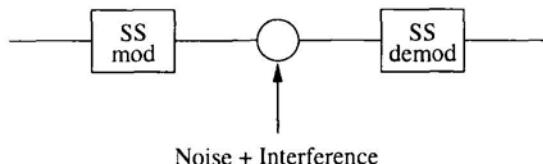
### 5.1 INTRODUCTION

Spread spectrum is a wireless communication technique that uses a transmission bandwidth, which is many times larger than the data rate. It has been used for military applications since the 40's. Its anti-jamming properties coupled with the fact that it allows the transmission of information hidden in the background noise makes it suitable for military purposes, where it is desired to make the transmitted signal hard to detect by the enemy.

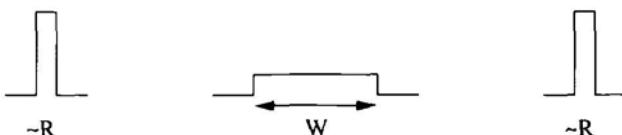
Recently, spread spectrum techniques have been applied to civilian applications for mobile telephone systems, where it is used as a multiple access communication technique that allows a large population of users to share a common channel. Furthermore, spread spectrum uses frequency diversity to combat the fading effects of the wireless multi-path environment.

The spread spectrum modulator spreads the information signal to a signal having a much larger bandwidth, as shown in Figure 5.1.b. The spreading process is done by a Pseudorandom Noise (PN) sequence, which is a high rate code having a much higher rate than that of the information signal. The transmitted signal appears to other users as noise. However, to the intended receiver, the wideband received signal is despread, using a PN sequence synchronized with that of the transmitter, such that the signal is despread to an identical replica of the transmitted signal, as shown in Figure 5.1.b.

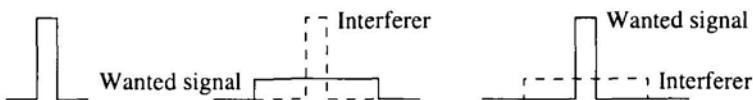
Figure 5.1.c shows the ability of the spread spectrum system to combat narrow-band interference. If the signal were transmitted with no spreading, using one of the narrow-band multiple access techniques such as TDMA or FDMA, and if the interference were to coincide on the transmitted signal, the receiver will not be able to properly detect that signal. However, in a spread spectrum system, the receiver's PN code which despreads the wanted signal,



(a)



(b)



(c)

*Figure 5.1.* The operation of spread spectrum systems and their ability to compact narrow band interference.

spreads the undesired signal (the interferer), hence reducing its power spectrum density, and enabling the use of a filter to remove the unwanted interference. Consequently, this mitigates the effect of the narrow-band interferer on the performance of the receiver.

In section 5.2, we present the basic equations governing the capacity of spread spectrum systems. There are three different spread spectrum modulation techniques, these are direct sequence spread spectrum, frequency hopping spread spectrum, and time hopping spread spectrum. These techniques are introduced in section 5.3. In section 5.4, we discuss the generation of pseudorandom noise (PN) sequences. The generation of PN sequences is an essential operation in any spread spectrum system. In section 5.5, we consider some implementation details of spread spectrum systems. These are synchronization, which is the process of achieving PN sequence synchronization between the receiver and the transmitter, tracking, which is the process of maintaining PN sequence synchronization as the mobile terminal changes its position, and power control,

which is the mobile terminal's ability to adjust its power so that the power received at the base station is the same for all terminals.

## 5.2 BASIC PRINCIPLES OF SPREAD SPECTRUM

The power of the spread spectrum system to suppress an interfering signal is called the processing gain ( $G$ ) of the spread spectrum system.  $G$  equals the ratio between the wideband signal bandwidth  $W$ , and that of the narrowband signal  $R$ :

$$G = \frac{W}{R} \quad (5.1)$$

Spread spectrum can be used as a multiple access technique, by allowing more than one user to transmit information over a common channel. Each user transmits its signal in the same frequency band and at the same time the other users transmit their signals. The receiver is required to distinguish the intended user from the unintended ones. Each channel is assigned a PN sequence distinguishing it from the other channels. The receiver is tuned to one PN sequence, thus despreading the desired channel signal, while the undesired channels pass through the despreader as noise.

Assume that there are  $N$  users, having equal power  $P$  at the receiver. The receiver despairs one of the transmitted signals. While, the remaining  $(N - 1)$  signals, which act as a source of interference to the desired signal, continue to have a bandwidth  $W$ .

The total interference power ( $I$ ) at the receiver is given by:

$$I = (N - 1)P \quad (5.2)$$

This power exists in the bandwidth  $W$ . Hence, the interference power spectral density ( $I_o$ ) is given by:

$$I_o = \frac{(N - 1)P}{W} \quad (5.3)$$

The desired signal has a bit-rate of  $R$  bits/sec and a power of  $P$  Watts. Hence, the desired signal bit energy is given by:

$$E_b = \frac{P}{R} \quad (5.4)$$

Therefore, the bit-energy-to-noise-density ratio, which determines the bit error rate performance of the receiver is given by:

$$\frac{E_b}{N_o} = \frac{W/R}{N - 1}$$

$$\approx \frac{W/R}{N} \quad (5.5)$$

$E_b/N_o$  is typically designed to be in the range of 3 - 9 dB. It depends on the modulation/demodulation scheme, the error-correcting code, channel impairments over the wireless link, and the bit error rate requirements. Equation 5.5, can be rewritten as:

$$N = \frac{W/R}{E_b/N_o} \quad (5.6)$$

$N$  is the total number of users that a spread spectrum system can accommodate while achieving a satisfactory bit error rate performance. There are two factors, which help increase the number of channels a spread spectrum system can support beyond that given by equation 5.6. These are the voice activity factor and the use of directional antennas.

During a telephone conversation, a person speaks for about 37.5% of the time [33], if we reduce the power transmitted over the channel to near zero during the periods of silence, the total number of channels the system can accommodate increases by the voice activity gain factor ( $G_V$ ). When a person speaks for 37.5% of the time,  $G_V = 2.67$ .

When using a  $v$ -sectored antenna. The coverage area is divided into  $v$  sectors, the total number of channels the system can accommodate increases by the antenna gain factor,  $G_A$ . Ideally,  $G_A = v$ . However, because the coverage area of each antenna is not sharply defined, there is overlap between the different sectors,  $G_A$  is less than  $v$ . For a three-sectored antenna,  $G_A = 2.4$  [34].

All cells in a spread spectrum system are allocated the same frequency band, i.e. the reuse factor is one. Hence, when calculating the total interference we must take into account the interference caused by the users of the adjacent cells. It was found that the interference caused by out of cell users is 60% of that caused by in cell users [34]. This interference ratio is denoted by  $f$ . Hence, the total number of channels a spread spectrum system supports is given by:

$$N = \frac{W/R}{E_b/N_o} \frac{G_V G_A}{1 + f} \quad (5.7)$$

### 5.3 SPREAD SPECTRUM TECHNIQUES

There are three spread spectrum modulation techniques: Direct Sequence Spread Spectrum (DSSS), Frequency Hopping Spread Spectrum (FHSS) and Time Hopping Spread Spectrum (THSS). In this section, we briefly describe each of these techniques.

### 5.3.1 Direct sequence spread spectrum

In the direct sequence spread spectrum system, the bipolar data is multiplied by a bipolar pseudorandom code generated by a pseudorandom number (PN) generator, as shown in Figure 5.2. Each bit of the pseudorandom code is known as a chip. The chip rate  $C$  is much larger than the input data rate  $R$  (to have effective spreading the chip rate needs to be at least four times the data rate). Thus, the output signal of the modulator  $m(t)$  has a much larger bandwidth than the input signal  $d(t)$ , and hence the spreading effect of the spread spectrum modulator is evident.

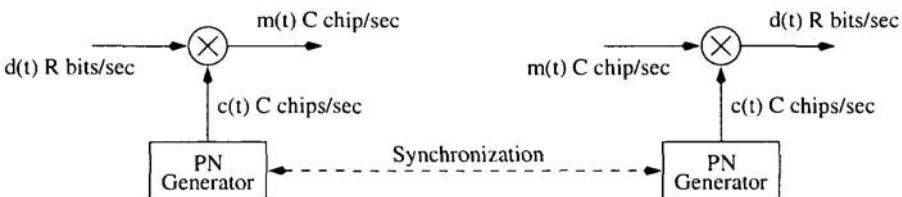


Figure 5.2. The direct sequence spread spectrum (DSSS) approach.

In the spread spectrum receiver, the wideband signal is despread as shown in Figure 5.2. A replica of the pseudorandom code is generated, this necessitates some type of synchronization between the PN generators of the receiver and transmitter. Figure 5.3 shows the waveforms involved in the direct sequence spread spectrum. The DSSS system experiences continuous but low rate random errors.

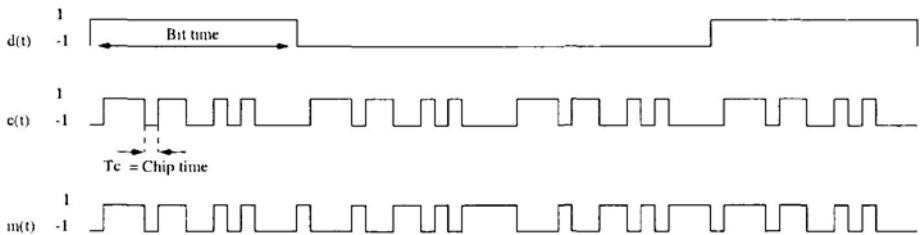


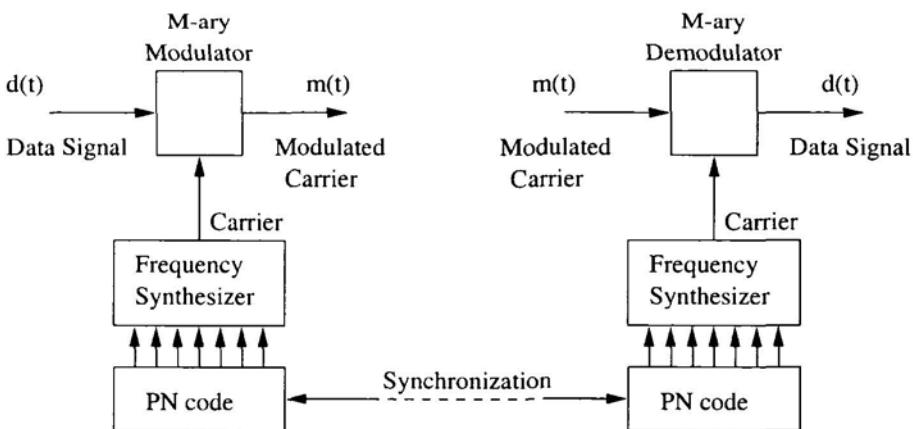
Figure 5.3. Waveforms of a direct sequence spread spectrum (DSSS) system.

### 5.3.2 Frequency hopping spread spectrum

In frequency hopping spread spectrum (FHSS), the carrier frequency changes (hops) according to a pseudorandom code sequence. The duration of each hop is known as the hop time  $T_h$ . The ratio between  $T_h$  and the input data symbol duration  $T_d$  determines the type of FHSS system:

- $\frac{T_h}{T_d} \gg 1 \longrightarrow$  Slow frequency hopping spread spectrum.  
There exists more than one data symbol per frequency hop.
- $\frac{T_h}{T_d} \ll 1 \longrightarrow$  Fast frequency hopping spread spectrum.  
There exists more than one frequency hop per data symbol.

Figure 5.4 shows an FHSS system, the PN generator generates a random code that controls the output frequency of the frequency synthesizer. The output signal of the frequency synthesizer is used as the carrier frequency to modulate the M-ary data. Hence, the output of the FHSS transmitter is a modulated M-ary signal having a carrier frequency dependent on the pseudorandom code generated by the PN generator.



*Figure 5.4. The frequency hopping spread spectrum (FHSS) approach.*

The FHSS system can be considered as a narrow band system that changes its carrier frequency over a much larger bandwidth. Hence, at a particular carrier frequency, the FHSS system can suffer interference and/or fading. This leads to occasional strong bursty errors. FHSS systems require frequency synthesizers having fine resolution with rapid settling and switching times.

### 5.3.3 Time hopping spread spectrum

In this system, the transmission time is divided into frames, which are further divided into slots. Each user is allowed to transmit during one and only one slot per frame. This system differs from a time division multiple access (TDMA) system in that the location of the time slot the user is allowed to transmit in varies randomly (pseudorandomly) from one frame to the next. Figure 5.5 shows a time diagram of the time hopping spread spectrum system.

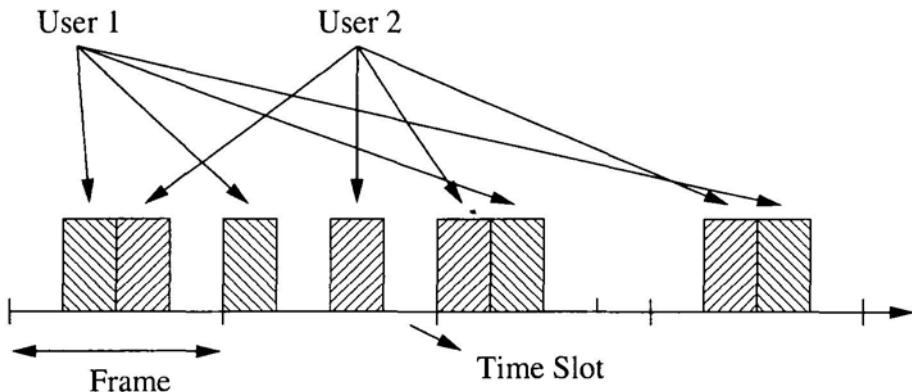


Figure 5.5. The time hopping spread spectrum (THSS) approach.

## 5.4 PSEUDORANDOM NOISE SEQUENCE

The generation of a binary independent sequence, known as the Bernoulli sequence is essential to the operation of a spread spectrum system. Because this random sequence has to be generated at both the receiver and transmitter, it needs to be deterministic, to allow its simultaneous generation at both the receiving and transmitting sides of the channel. This sequence is known as the pseudorandom noise (PN) sequence.

The PN sequence has two functions in a direct sequence spread spectrum system. First, it spreads the bandwidth of the modulated signal to the larger transmission bandwidth. Second, it distinguishes the different users.

Ideally, a random binary sequence needs to satisfy the following requirements:

1. Half the bits are zeros and half of them are ones.
2. Half the run lengths of the ones or zeros should be of one-bit length, quarter of them should be of two-bit length, one-eighth of them should be of three-bit length, and so on.
3. When shifted by a nonzero number of bits that is not a multiple of the period, the number of agreements and disagreements between the original sequence and the shifted sequence is the same.

The PN sequence is a periodic binary sequence that is generated using a linear feedback shift register, such as that shown in Figure 5.6. The shift register consists of a cascade of memory elements (delay blocks). The outputs of the memory elements are logically combined to produce the input to the first stage. An  $n$ -stage shift register can be in any of  $2^n$  states, as determined by the  $n$ -bit vector it stores. One of these states is the zero state, where all the  $n$ -bits

are zeros. A linear feedback shift register that generates a PN sequence must go cyclically through all of the  $(2^n - 1)$  non-zero states. Hence, if the shift register contains  $n$ -stages, the generated PN sequence should have a period  $2^n - 1$ . This sequence is referred to as a maximal length sequence, and it has the following properties:

1. There are  $2^{n-1}$  ones and  $2^{n-1} - 1$  zeros.
2. There are  $2^{n-3}$  one-bit runs of ones and zeros.  
There are  $2^{n-4}$  two-bit runs of ones and zeros.

.....  
There is one  $(n - 2)$ -bit run of ones and zeros.

There is one  $(n - 1)$ -bit run of zeros.

There is one  $n$ -bit run of ones.

Table 5.1 shows the number of runs for the PN sequence generated using a 5-stage shift register.

3. During each period, the shift register passes cyclically through the  $2^n - 1$  non-zero states.

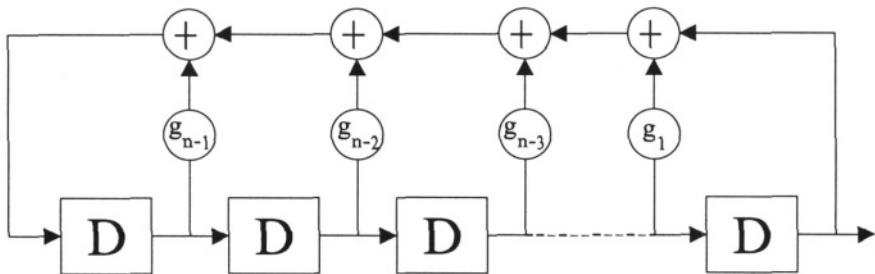


Figure 5.6. The Fibonacci Linear Feedback Shift Register.

The linear feedback shift register of Figure 5.6, is known as Fibonacci shift register. An alternative shift register, that can be used to generate the same PN sequence is the Galois shift register, which is shown in Figure 5.7. The shift register coefficients,  $g_i$  and  $h_i$  are related by the following equation:

$$g_i = h_{n-i} \quad (5.8)$$

Where,  $n = 1 \dots (n - 1)$ .

One advantage of the Galois form over the Fibonacci form, is that the former avoids the long chain of XORs, which makes it faster than the latter.

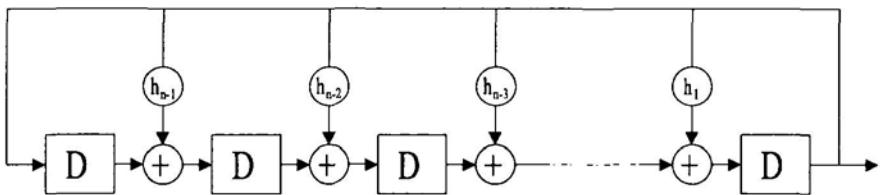


Figure 5.7. The Galois Linear Feedback Shift Register.

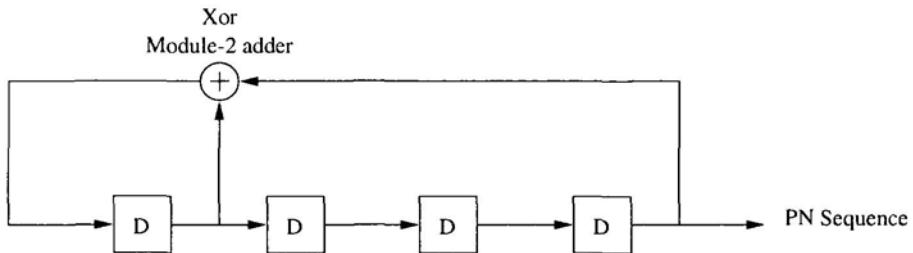


Figure 5.8. A four stage linear feedback shift register used to generate PN sequences.

Table 5.1. Number of one and zero runs of length  $n$  of a PN sequence generated using a 5-stage linear feedback shift register.

Run length	Ones	Zeros	Total number of bits
1	4	4	8
2	2	2	8
3	1	1	6
4		1	4
5	1		5
Total number of bits			31

Figure 5.8 shows a four stage Fibonacci linear Feedback Shift Register. The input of the first stage of the linear feedback shift register of Figure 5.8 can be expressed as:

$$Y_n = Y_{n-1} + Y_{n-4} \quad (5.9)$$

The addition operator in equation 5.9 is module-2 addition. Knowing that in module-2 addition  $a + a = 0$ , equation 5.9 can be rewritten as:

$$Y_n + Y_{n-1} + Y_{n-4} = 0 \quad (5.10)$$

$D$  is the delay operator, accordingly:

$$(1 + D + D^4)Y = 0 \quad (5.11)$$

Where,  $(1 + D + D^4)$  is the characteristic polynomial of the linear feedback shift register shown in Figure 5.8. In general, each linear feedback shift register has its own characteristic equation. For an  $n$ -stage Fibonacci shift register as that shown in Figure 5.6, the  $n$ -degree characteristic polynomial is given by:

$$1 + g_{n-1}D + g_{n-2}D^2 + \cdots + g_1D^{n-1} + D^n \quad (5.12)$$

Where,  $g_1, g_2 \dots g_{n-1}$  of equation 5.12, are as shown in Figure 5.6. In order to generate a maximal length sequence having a period  $(2^n - 1)$ , the polynomial of equation 5.12 must be an irreducible primitive polynomial [35]. An irreducible polynomial is a polynomial that can't be factorized. For example,  $x^2 + 1$  is not an irreducible polynomial because:

$$x^2 + 1 = (x + 1)(x + 1) \quad (5.13)$$

On the other hand  $x^2 + x + 1$  is an irreducible polynomial, i.e. it can't be factorized. An irreducible polynomial of degree  $n$  is said to be primitive if the smallest degree  $d$  of a polynomial having the form  $x^d + 1$ , it is a factor of, is:

$$d = 2^n - 1 \quad (5.14)$$

For an  $n$ th-degree polynomial, corresponding to an  $n$ -stage shift register, the number of primitive polynomials is given by [35]:

$$\frac{2^n - 1}{n} \prod_{i=1}^k \frac{P_i - 1}{P_i} \quad (5.15)$$

Where,  $P_i$  is the prime decomposition of  $2^n - 1$ . Table 5.2 gives the number of primitive polynomials of degree  $n$ , along with an example polynomial for each degree.

Table 5.2. Number of primitive polynomials of degree  $N$ ,  $N = 2 - 6$ .

$n$	$2^n - 1$	Number of Primitive Polys	Example
2	3	$\frac{3}{2} \cdot \frac{2}{3} = 1$	$1 + D + D^2$
3	7	$\frac{7}{3} \cdot \frac{3}{7} = 2$	$1 + D + D^3$
4	$15 = 3 \times 5$	$\frac{15}{4} \cdot \frac{2}{3} \cdot \frac{4}{5} = 2$	$1 + D + D^4$
5	31	$\frac{31}{5} \cdot \frac{30}{31} = 6$	$1 + D^2 + D^5$
6	$63 = 3^2 \times 7$	$\frac{63}{6} \cdot \frac{2}{3} \cdot \frac{6}{7} = 6$	$1 + D + D^6$

**Example 5.1.** For the shift register of Figure 5.8, assume that the initial state of the register is 1 0 0 0. Find the output sequence and the state of the register during the next 16 clock cycles.

**Solution.** Table 5.3 shows the register state, in both binary and decimal representations, as it progresses from one cycle to the next, as well as the output sequence. Notice that, after 15 clock cycles, the shift register returns back to its initial state, and the cycle repeats itself. Also notice that the shift register cycles through all the 15 nonzero states. Hence, the generated PN sequence is a maximal length sequence.

Table 5.3. The output sequence of the linear feedback shift register of Figure 5.8.

Clock cycle	State		Output
	Binary	Decimal	
0	1 0 0 0	8	0
1	1 1 0 0	12	0
2	1 1 1 0	14	0
3	1 1 1 1	15	1
4	0 1 1 1	7	1
5	1 0 1 1	11	1
6	0 1 0 1	5	1
7	1 0 1 0	10	0
8	1 1 0 1	13	1
9	0 1 1 0	6	0
10	0 0 1 1	3	1
11	1 0 0 1	9	1
12	0 1 0 0	4	0
13	0 0 1 0	2	0
14	0 0 0 1	1	1
15	1 0 0 0	8	0
16	1 1 0 0	12	0

So far, we have represented each symbol of the PN sequence as {0,1}. The addition operation performed to generate the spread spectrum signal from the PN sequence and the data stream is module-2 addition:

$$(\text{Spread Spectrum Signal}) = (\text{Data Stream}) \oplus (\text{PN Sequence}) \quad (5.16)$$

The module two addition operation satisfies the following set of equations:

$$0 \oplus 0 = 0 \qquad \qquad \qquad 1 \oplus 0 = 1$$

$$0 \oplus 1 = 1 \qquad \qquad \qquad 1 \oplus 1 = 0$$

Alternatively, each symbol of the PN sequence and data stream can be represented as {1, -1}. Where, 1 corresponds to logic 0 and -1 corresponds to logic 1. The module-2 addition operation is replaced by multiplication:

$$(\text{Spread Spectrum Signal}) = (\text{Data Stream})(\text{PN Sequence}) \quad (5.17)$$

The multiplication operation satisfies the following set of equations:

$$\begin{array}{ll} 1 \cdot 1 = 1 & -1 \cdot 1 = -1 \\ 1 \cdot -1 = -1 & -1 \cdot -1 = 1 \end{array}$$

## 5.5 PRACTICAL CONSIDERATIONS IN SPREAD SPECTRUM SYSTEMS

### 5.5.1 Synchronization

Timing synchronization is the process of determining the timing of the receiver's PN sequence to match that of the transmitter, such that if the received chip sequence is correlated with the receiver's PN sequence the transmitted data is recovered at the receiver.

Synchronization involves testing all the likely hypotheses for a correct value of the timing parameter [34]. This testing can proceed in parallel or serially. The former requires more hardware resources than the latter, while the latter can take substantially longer time to complete. Practically, a combination of the two is implemented in such a way that a number of devices proceed in parallel, with each assigned a subset of the total hypotheses.

Another timing synchronization strategy that can be employed is to have two passes through the hypotheses, the first pass is a short pass that eliminates the least likely hypotheses, while the second pass is a longer pass that carefully considers the likely hypotheses to come up with the correct value of the timing parameter.

Consider a baseband direct sequence spread spectrum system with a spreading factor  $M$ . Let  $S$  be the received signal power, hence the received bit energy is given by:

$$E_b = ST_b \quad (5.18)$$

Where,  $T_b$  is the bit duration. The noise power  $N$  is related to the noise power spectral density  $N_o$  by the following equation:

$$N = N_o \text{BW} \quad (5.19)$$

Where, BW is the bandwidth of the baseband-spread signal, and it is related to the bit duration ( $T_b$ ) by the following equation:

$$\text{BW} = \frac{M}{2T_b} \quad (5.20)$$

Therefore, in a baseband direct sequence spread spectrum system,  $E_b/N_o$  is related to the signal-to-noise ratio (SNR) by the following equation:

$$\frac{E_b}{N_o} = \frac{M}{2} \text{SNR} \quad (5.21)$$

Assume a spread spectrum system as shown in Figure 5.9. The data on the transmitter side is multiplied by the PN sequence to generate the transmitted chips, which in turn are filtered by a square-root raised cosine filter. The filtered signal is then transmitted across the channel, where additive white Gaussian noise, having a power spectrum density  $N_o$ , is added to it.

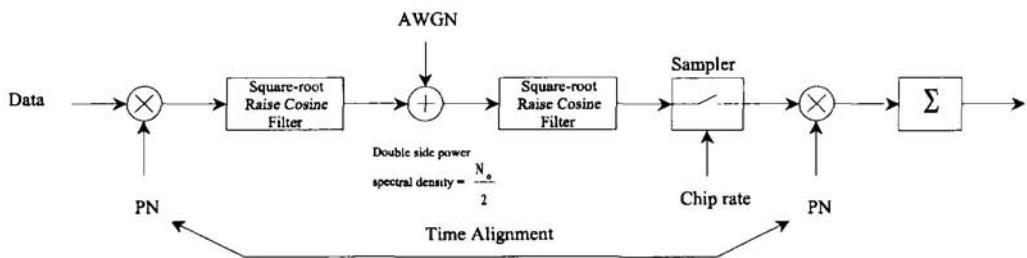


Figure 5.9. A direct sequence spread spectrum (DSSS) system.

On the receiver side, the received signal is filtered by a square root raised cosine filter. The filtered signal is then sampled at the chip rate and correlated with the receiver's PN sequence to recover the data.

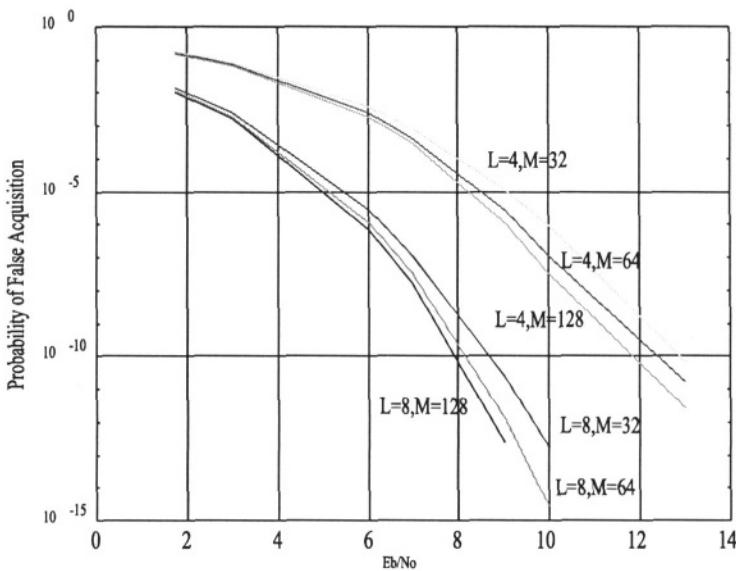
The synchronization process involves determining the correct timing of the receiver's PN sequence to be aligned with that of the transmitter, after taking into account the channel delay. The pilot channel is used in the synchronization process. The transmitted data on the pilot channel is always logic "0".

To acquire the PN sequence, the received chip stream is correlated with different offsets of the PN sequence. The correlation that gives the largest output corresponds to the acquired PN sequence.

A false acquisition occurs when the acquired PN sequence doesn't have the correct timing offset. In the following analysis we attempt to calculate the probability of false acquisition when the sampling time is aligned to a chip.

Assume an AWGN environment, having a noise power spectrum density  $N_o$ . Furthermore, assume that the spreading code is a real valued PN sequence having a chip rate  $M$  times the data rate. The synchronization search window searches  $K$  offsets each separated by one chip, such that the sampling points are aligned to a chip.  $L$  is the number of symbols correlated per offset. Assume that the output of the receiver's square-root raised cosine filter is normalized to be  $\pm 1$  when sampled at the ISI-free point, and in the absence of noise. Hence, the output of the correlator (having  $L \times M$  chip window) when the PN sequence has the correct timing offset has a mean given by:

$$\mu_x = ML \quad (5.22)$$



*Figure 5.10.* Probability of false acquisition versus  $E_b/N_0$ . The synchronization search window is 30 offsets, each separated by one chip. The samples are perfectly aligned to the chips.

In the presence of noise, the variance is given by:

$$\sigma_x^2 = \frac{ML}{\text{SNR}} \quad (5.23)$$

The correlator output is represented by a Gaussian random variable  $X$ , with mean and variance as given by equations 5.22 and 5.23 respectively. When the timing of the PN sequence is off by an integer multiple of chips, the output of the correlator is represented by a random variable  $Y$  that has a zero mean and a variance given by:

$$\sigma_y^2 = ML + \frac{ML}{\text{SNR}} \quad (5.24)$$

Figure 5.10 shows the probability of false acquisition.

### 5.5.2 Tracking

Synchronization is only capable of synchronizing the PN sequences of the receiver and transmitter to within a fraction of a chip. Tracking is needed to fine tune the timing of the PN sequence. Furthermore, tracking is needed

to maintain the PN sequences synchronization as the mobile terminal moves around.

A common tracking technique is the so-called early-late tracking. Figure 5.11 displays a spread spectrum system employing early late tracking. The early-late tracker consists of three correlators, the early correlator, the late correlator and the on-time correlator. Typically, the on-time correlator samples the output of the receiver at its peak. The early correlator samples the output of the receiver filter  $\Delta$  seconds before the on-time correlator. While the late correlator samples the output of the receiver filter  $\Delta$  seconds after the on-time correlator.

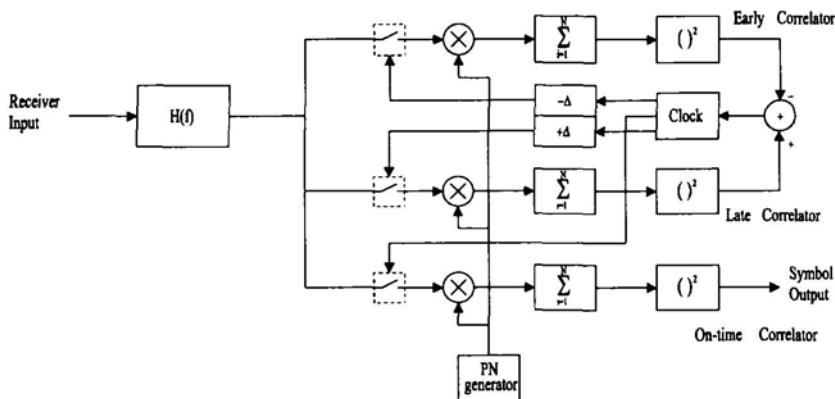


Figure 5.11. Early Late Tracker.

The peak of the correlator's output occurs when the sampler samples the received signal at the zero-IS point, and the receiver's PN-sequence is properly synchronized to the PN-sequence generating the received signal. As the timing of the sampler starts to advance or retard from its correct (optimum) value, the output signal from the correlator starts to drop. If the timing of the sampler is off by one full chip or more from its correct value, the output of the correlator is zero (random noise).

Figure 5.12.a shows the output of the correlator versus the sampling time of the sampler. Notice that the correlator's output is symmetric around its peak value. The on-time correlator samples the received signal at the correct (optimum) time, which produces the peak output from the correlator. The early and late correlators produce the same output due to the symmetry of the function. Hence, the difference between the output of the early and late correlators is zero, and no correction is applied to the clock of the sampler. This case is shown in Figure 5.12.a.

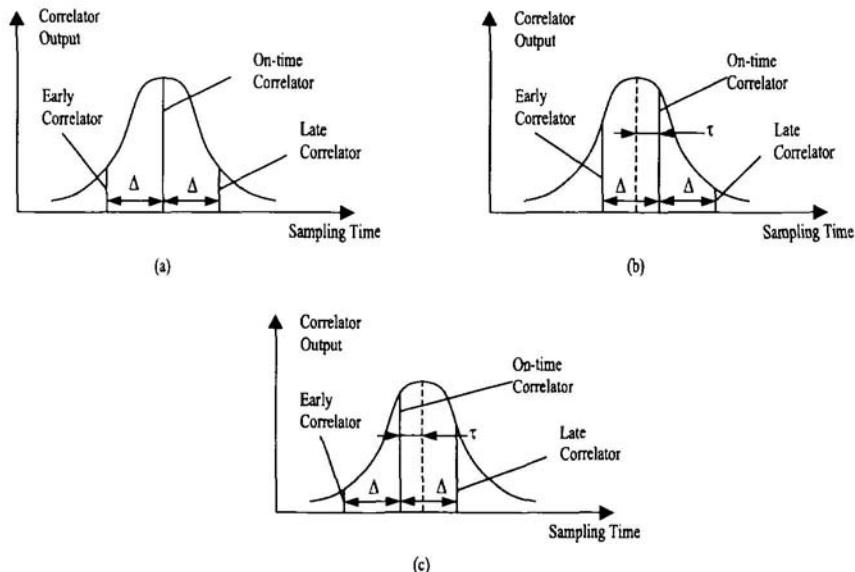


Figure 5.12. On-time, early and late correlator outputs: (a) On-time correlator samples received signal at the optimum time (b) Sampler of on-time correlator retarded  $\tau$  seconds. (c) Sampler of on-time correlator advanced  $\tau$  seconds.

If the on-time correlator samples the received signal  $\tau$  seconds after the correct (optimum) time, as shown in Figure 5.12.b, the output of the early correlator is larger than that of the late correlator. Hence, the difference between the outputs of the early and late correlators is negative, this speeds up the clock of the sampler so as to reduce  $\tau$  and push the on-time correlator's output to its peak value.

If the on-time correlator samples the received signal  $\tau$  seconds before the correct (optimum) time, as shown in Figure 5.12.C, the output of the late correlator is larger than that of the early correlator. Hence, the difference between the outputs of the early and late correlators is positive, this slows down the clock of the sampler so as to reduce  $\tau$  and push the on-time correlator's output to its peak value.

### 5.5.3 Power Control

The near far problem, where strong signals interfere with the detection of weak signals, can cause substantial degradation in the performance of a CDMA system. This problem typically appears in the reverse channel (up link), where the mobile terminals are randomly scattered inside a cell. If all the mobile

terminals were allowed to transmit with the same power level, those closer to the base station, will have stronger power level than those close to the edge of the cell. This will in turn impact the system's capacity.

To mitigate the effect of the near far problem, power control is employed, such that power received at the base station from any mobile terminal is almost equal to that of any other mobile user, in order to maximize the system's capacity. Power control is required to have a wide dynamic range of 80 to 100 dB, in order to accommodate the variations in the received power due to varying distances from the base station, as well as shadowing effects.

There are two ways to achieve power control. The first, is the "open loop" technique, which is solely under the control of the mobile terminal. The mobile terminal determines the strength of the signal on the forward link, this is used to calculate the propagation loss between base station and the mobile terminal and accordingly the power that needs to be transmitted by the mobile terminal to achieve a certain signal to interference ratio ( $E_b/I_o$ ) at the receiver of the base station.

However, open loop power control is not effective enough, the reverse link received power at the base station can vary by a few decibels. This is because the propagation loss on the reverse and forward links are not identical, because the forward and reverse link have different frequencies. Furthermore, diversity reception might be in use on the reverse link and not on the forward link.

To overcome this, "closed loop" power control is needed. In this case power control is handled at the base station. When the base station determines that  $E_b/I_o$  of a particular user is above or below certain thresholds, it transmits an instruction to the mobile terminal to decrease or increase its power by  $\Delta$  dB respectively.

In IS-95 the base station requests the mobile terminal to change its transmit power in 1 dB increments once every 1.25 msec through an 800-bps power control channel that is embedded into the traffic channel.

## Chapter 6

# RECEIVER ARCHITECTURES

### 6.1 INTRODUCTION

In this chapter we consider the design of a wireless receiver. We start by considering the phenomena that impact the performance of the receiver, such as noise, which is considered in section 6.2, and intermodulation distortion, which is considered in section 6.3. Then we present some commonly used receiver architectures, such as the superheterodyne receiver in section 6.4, the homodyne receiver in section 6.5, and software radio, which is a fully digital and programmable transceiver in section 6.6.

The performance of a wireless receiver is best described in terms of its figures of merit, these include:

**Noise figure.** This is the ratio between the signal-to-noise ratio at the input of the receiver to that at the output of the receiver. The noise figure can also be described in terms of the noise temperature.

**Linear dynamic range.** This is the input signal range between the receiver's threshold, where the signal-to-noise ratio is 0 dB, to the linearity limit, where the gain of the receiver drops by 1 dB from its small signal value.

**Spurious-free dynamic range.** This is the input signal range between the receiver's threshold, and input signal level where second or third-order non-linear effects equal the threshold level. The spur-free dynamic range is usually less than the linear dynamic range.

### 6.2 NOISE FIGURE

Any receiver adds noise to the received signal. The amount of noise added by the receiver is determined by a receiver parameter known as the noise figure

(NF). In the case of an ideal receiver, which adds no noise to the incoming signal, the noise figure is one. Practically, the noise figure is larger than one.

Associated with each noise figure (NF) is an effective noise temperature ( $T_e$ ), which is related to NF as explained in the following. Assume that the receiver operates at room temperature,  $T_o = 290K$  ( $17^\circ C$ ). Then, the input noise to the receiver, referring to Figure 6.1, is given by:

$$N_i = kT_o B \quad (6.1)$$

Where,

$k$  is Boltzman's constant  $= 1.38 \times 10^{-23}$  Joule/K.

$B$  is the bandwidth of the input signal.

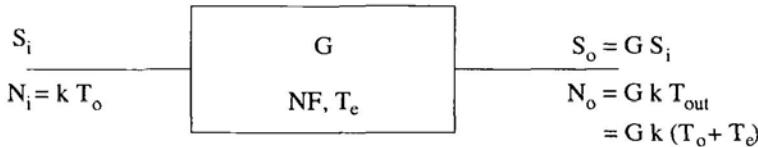


Figure 6.1. Noise figure in a one stage receiver.

Assume that the signal-to-noise ratio at the input of the receiver is  $\text{SNR}_i$ , and that at the output of the receiver is  $\text{SNR}_o$ . The noise figure of the receiver is defined as the ratio between the input and output signal-to-noise ratios.

$$NF = \frac{\text{SNR}_i}{\text{SNR}_o} \quad (6.2)$$

If the receiver has a gain  $G$ , the noise at the output of the receiver is related to that at the input of the receiver by the following equation:

$$\begin{aligned} N_o &= N_i G NF \\ &= G k T_{out} \end{aligned} \quad (6.3)$$

$T_{out}$  is the noise temperature at the output of the receiver. Associated with each receiver is a temperature  $T_e$  known as the effective noise temperature of the receiver.  $T_e$  represents the temperature of the extra noise added by the receiver, and is related to the noise temperatures, at the input and output of the receiver, by the following equation:

$$T_{out} = T_o + T_e \quad (6.4)$$

From equations 6.1-6.4, NF and  $T_e$  are related by the following expression:

$$\text{NF} = 1 + \frac{T_e}{T_o} \quad (6.5)$$

Consider the case where the receiver consists of two stages in cascade as shown in Figure 6.2. The first stage has a power gain  $G_1$ , a noise figure  $\text{NF}_1$ , and a noise temperature  $T_{e1}$ . The second stage has a power gain  $G_2$ , a noise figure  $\text{NF}_2$ , and a noise temperature  $T_{e2}$ . The objective of the following analysis is to determine the overall noise figure of the receiver in terms of the parameters of its constituent stages.

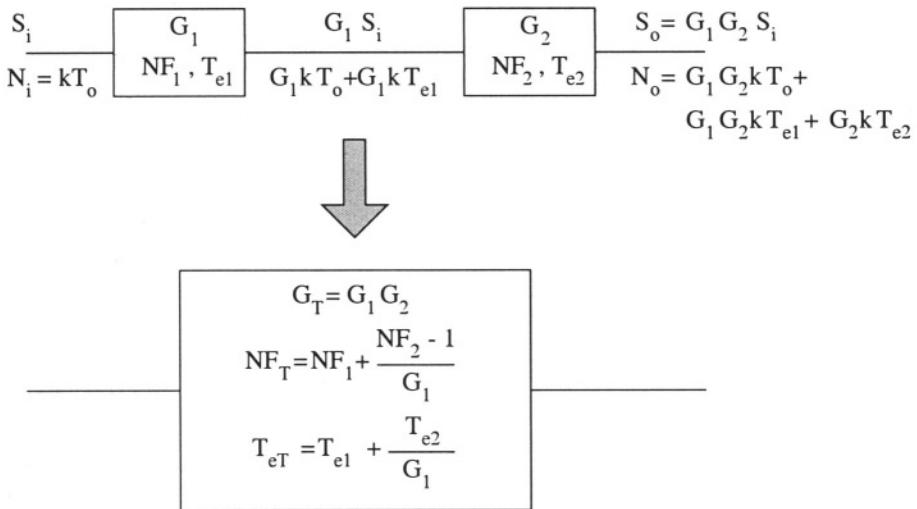


Figure 6.2. The noise figure of a two stage receiver.

The noise at the input to the receiver is given by equation 6.1. The noise at the output of the first stage  $N_1$  is given by:

$$N_1 = G_1 k(T_o + T_{e1}) \quad (6.6)$$

The noise at the output of the receiver (after the second stage) is given by:

$$\begin{aligned} N_o &= G_2 N_1 + G_2 k T_{e2} \\ &= G_1 G_2 k T_o \left[ 1 + \frac{T_{e1}}{T_o} + \frac{T_{e2}}{G_1 T_o} \right] \end{aligned} \quad (6.7)$$

The signal at the output of the receiver is related to that at the input by the following equation:

$$S_o = G_1 G_2 S_i \quad (6.8)$$

Therefore, substituting for the signal-to-noise ratios in equation 6.2, we get the overall noise figure of the two stage receiver, which is given by:

$$\begin{aligned} \text{NF} &= 1 + \frac{T_{e1}}{T_o} + \frac{T_{e2}}{G_1 T_o} \\ &= \text{NF}_1 + \frac{\text{NF}_2 - 1}{G_1} \end{aligned} \quad (6.9)$$

The effective noise temperature of the two-stage amplifier is given by:

$$T_e = T_{e1} + \frac{T_{e2}}{G_1} \quad (6.10)$$

Notice that, the noise figure of the first stage  $\text{NF}_1$  has more influence in determining the overall noise figure of the receiver, than the noise figure of the second stage  $\text{NF}_2$ . Furthermore, the larger the gain  $G_1$  of the first stage, the smaller the effect of the noise figure  $\text{NF}_2$  of the second- stage on the overall noise figure of the receiver.

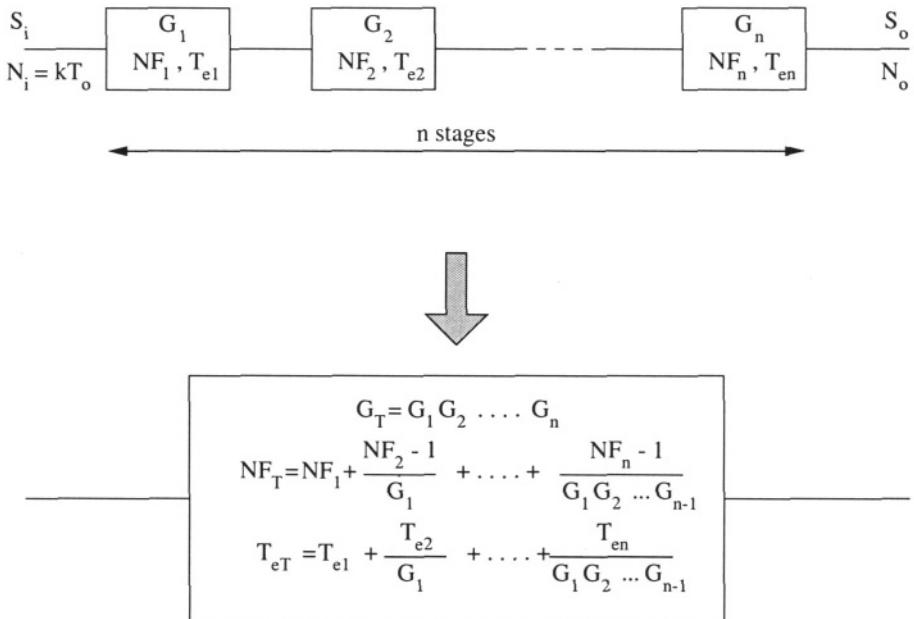


Figure 6.3. The noise figure of an  $n$ -stage receiver.

Extending these results to an  $n$ -stage receiver as that of Figure 6.3, we get the following expression for the noise figure of the receiver:

$$NF = NF_1 + \frac{NF_2 - 1}{G_1} + \frac{NF_3 - 1}{G_1 G_2} + \dots + \frac{NF_n - 1}{G_1 G_2 \dots G_{n-1}} \quad (6.11)$$

The effective noise temperature of an  $n$ -stage amplifier is given by:

$$T_e = T_{e1} + \frac{T_{e2}}{G_1} + \frac{T_{e3}}{G_1 G_2} + \dots + \frac{T_{en}}{G_1 G_2 \dots G_{n-1}} \quad (6.12)$$

**Example.** Consider the receiver shown in Figure 6.4. Find the noise figure of the receiver, and the equivalent noise temperature.

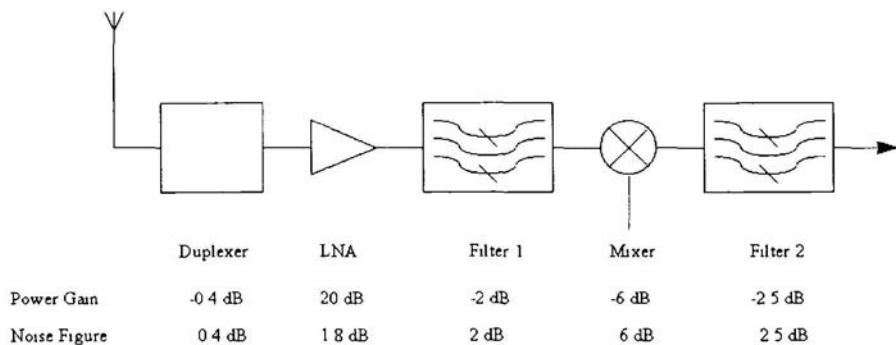


Figure 6.4. Receiver used in the calculation of the noise figure.

**Solution.** The values of the gain and noise figure given in Figure 6.4 are in dBs, these are converted to absolute values as shown in Table 6.1. The noise figure is calculated by substituting the values of Table 6.1 into equation 6.11. The noise figure is:

$$\begin{aligned} NF &= 1.77 \\ &= 2.48 \text{dB} \end{aligned}$$

The effective noise temperature of the receiver is calculated by substituting the values of Table 6.1 into equation 6.12. The effective noise temperature is:

$$T_e = 223.78 \text{ K}$$

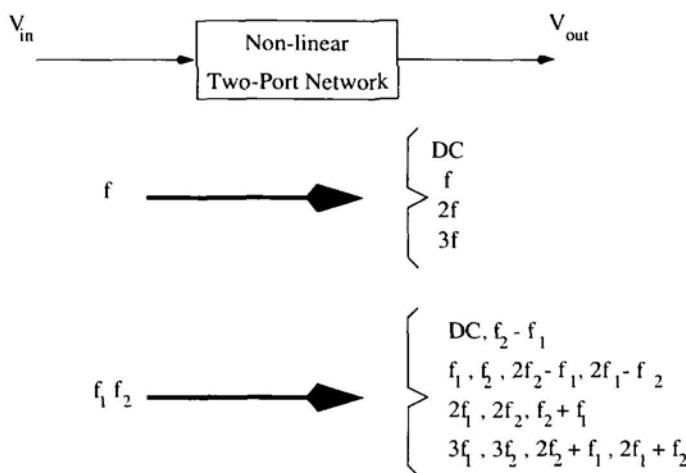
*Table 6.1.* Power gain, noise figure, effective noise temperature for each stage of the receiver shown in Figure 6.4.

Device	Power Gain	Noise Figure	Noise Temperature
Duplexer	$G_1 = 0.912$	$NF_1 = 1.096$	$T_{e1} = 27.84 \text{ K}$
LNA	$G_2 = 100$	$NF_2 = 1.514$	$T_{e2} = 149.06 \text{ K}$
Filter1	$G_3 = 0.631$	$NF_3 = 1.584$	$T_{e3} = 169.36 \text{ K}$
Mixer	$G_4 = 0.251$	$NF_4 = 3.981$	$T_{e4} = 864.49 \text{ K}$
Filter2	$G_5 = 0.562$	$NF_5 = 1.778$	$T_{e5} = 225.62 \text{ K}$

### 6.3 INTERMODULATION DISTORTION

A typical receiver exhibits non-linear effects. To illustrate the effect of non-linearity on the operation of a receiver consider a two port non-linear network as that shown in Figure 6.5. Assume that the two port network has only second and third order non-linearities. Thus, its transfer function is given by:

$$v_{\text{out}} = a_1 v_{\text{in}} + a_2 v_{\text{in}}^2 + a_3 v_{\text{in}}^3 \quad (6.13)$$



*Figure 6.5.* A two port non-linear network to illustrate non-linear distortion.

Assume that a sinusoidal wave having a frequency  $f$  is injected into the two port network:

$$v_{\text{in}} = V \cos 2\pi f t \quad (6.14)$$

The linear term of equation 6.13 ( $a_1 v_{\text{in}}$ ) contributes to the output signal by:

$$^1 v_{\text{out}} = V a_1 \cos 2\pi f t \quad (6.15)$$

The second-order non-linear term of equation 6.13 ( $a_2 v_{\text{in}}^2$ ) contributes to the output signal by:

$$^2 v_{\text{out}} = \frac{V^2 a_2}{2} [1 + \cos 2\pi(2f)t] \quad (6.16)$$

This term has a DC component and a second-order harmonic. The third-order non-linear term of equation 6.13 ( $a_3 v_{\text{in}}^3$ ) contributes to the output signal by:

$$^3 v_{\text{out}} = \frac{V^3 a_3}{4} [3 \cos 2\pi f t + \cos 2\pi(3f)t] \quad (6.17)$$

This term has a fundamental component and a third-order order harmonic. Thus, the output of the non-linear two-port network shown in Figure 6.5, when a sinusoidal signal having a frequency  $f$  is injected into it, contains a DC term, a second-order harmonic at frequency  $2f$ , a third-order harmonic at frequency  $3f$ , in addition to the fundamental component at frequency  $f$ . The effect of the DC component and the harmonics at  $2f$ ,  $3f$ , ... can be alleviated by filtering them out. However, the real detrimental effect of non-linearity is due to intermodulation distortion.

Consider two sinusoidal waveforms with frequencies  $f_1$  and  $f_2$  are injected into a non-linear two-port network such as that shown in Figure 6.5:

$$v_{\text{in}} = V_1 \sin 2\pi f_1 t + V_2 \sin 2\pi f_2 t \quad (6.18)$$

Where,  $f_2$  is slightly larger than  $f_1$ . The linear term of equation 6.13 ( $a_1 v_{\text{in}}$ ) contributes to the output signal by:

$$^1 v_{\text{out}} = V_1 a_1 \cos 2\pi f_1 t + V_2 a_2 \cos 2\pi f_2 t \quad (6.19)$$

The second-order non-linear term of equation 6.13 ( $a_2 v_{\text{in}}^2$ ) contributes to the output signal by:

$$^2 v_{\text{out}} = \underbrace{V_1^2 a_2 \cos^2 2\pi f_1 t + V_2^2 a_2 \cos^2 2\pi f_2 t}_{\text{DC + Second order harmonic}} +$$

$$\frac{2V_1V_2a_2 \cos 2\pi f_1 t \cos 2\pi f_2 t}{V_1V_2a_2[\cos 2\pi(f_1+f_2)t + \cos 2\pi(f_2-f_1)t]} \quad (6.20)$$

This term has a DC component, two second-order harmonics at  $2f_1$  and  $2f_2$  and two intermodulation distortion components at  $f_1 + f_2$  and  $f_2 - f_1$ . The third-order non-linear term of equation 6.13 contributes to the output signal by:

$$\begin{aligned} {}^3v_{\text{out}} = & \underbrace{V_1^3 a_3 \cos^3 2\pi f_1 t + V_2^3 a_3 \cos^3 2\pi f_2 t}_{\text{Fundamental component + Third order harmonic}} + \\ & \underbrace{3V_1^2 V_2 a_3 \cos^2 2\pi f_1 t \cos 2\pi f_2 t}_{\frac{3}{4}\{V_1^2 V_2 a_3 [2 \cos 2\pi f_2 t + \cos 2\pi(2f_1+f_2)t + \cos 2\pi(2f_1-f_2)t]\}} \\ & \underbrace{3V_1 V_2^2 a_3 \cos 2\pi f_1 t \cos^2 2\pi f_2 t}_{\frac{3}{4}\{V_1^2 V_2 a_3 [2 \cos 2\pi f_1 t + \cos 2\pi(f_1+2f_2)t + \cos 2\pi(2f_2-f_1)t]\}} \end{aligned} \quad (6.21)$$

This term has the fundamental components  $f_1$  and  $f_2$ , two third-order harmonics  $3f_1$  and  $3f_2$ , and four intermodulation distortion components at  $2f_1 + f_2$ ,  $2f_1 - f_2$ ,  $f_1 + 2f_2$  and  $2f_2 - f_1$ . If the non-linear two-port network had a higher degree of non-linearity, there would have been more spectral components at the output.

The spectral components at the output of a third-order non-linear two-port network are shown in Figure 6.6. The spectral components  $f_1 + f_2$ ,  $f_2 - f_1$ ,  $2f_1 + f_2$ ,  $2f_1 - f_2$ ,  $2f_2 + f_1$ , and  $2f_2 - f_1$  are known as the intermodulation distortion components.

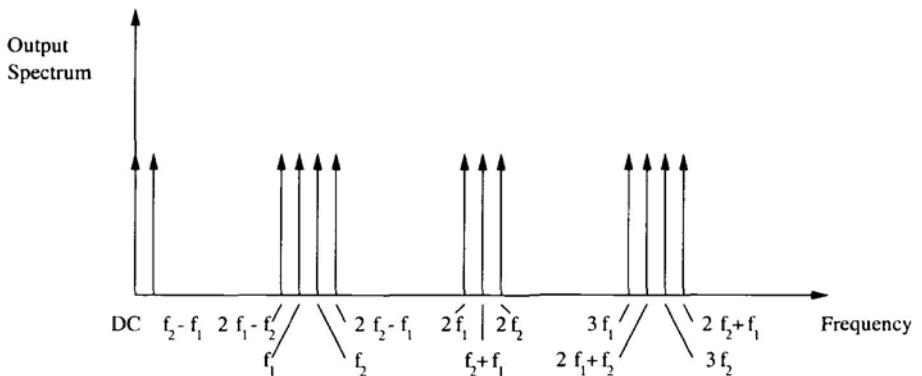


Figure 6.6. The frequency spectrum of the output of a third-order non-linear two-port network.

Of all the intermodulation distortion components in the frequency spectrum of the output signal,  $2f_2 - f_1$  and  $2f_1 - 2f_2$  are the only two components

that can't be filtered out, and hence interfere with the wanted signal. Since,  $2f_2 - f_1$  and  $2f_1 - f_2$  are inside the band of interest, they are usually referred to as the in-band intermodulation distortion components.

According to equations 6.16 and 6.20, the second-order non-linear terms are proportional to the square of the input voltage, i.e. every 1 dB change in the input signal, causes the second-order non-linear terms to change by 2 dB. Similarly, according to equations 6.17 and 6.21, the third-order non-linear terms are proportional to the cube of the input voltage, i.e. every 1 dB change in the input signal, causes the third-order non-linear terms to change by 3 dB.

Figure 6.7 is a plot of the output signal power showing, the linear term, the second-order non-linear term and the third-order non-linear term, versus the input voltage on a log-log scale. The linear term intersects with the second-order non-linear term at the second-order intercept point. This corresponds to the input signal-level at which the output signal power due to the linear term of equation 6.13 is equal to that due to the second-order non-linear term of that equation. Similarly, the third-order intercept point is the intersection of the plot corresponding to the linear term of equation 6.13, with that of the third-order non-linear term. This point corresponds to the input signal level at which the output signal power due to the linear term of equation 6.13 is equal to that of the third-order non-linear term.

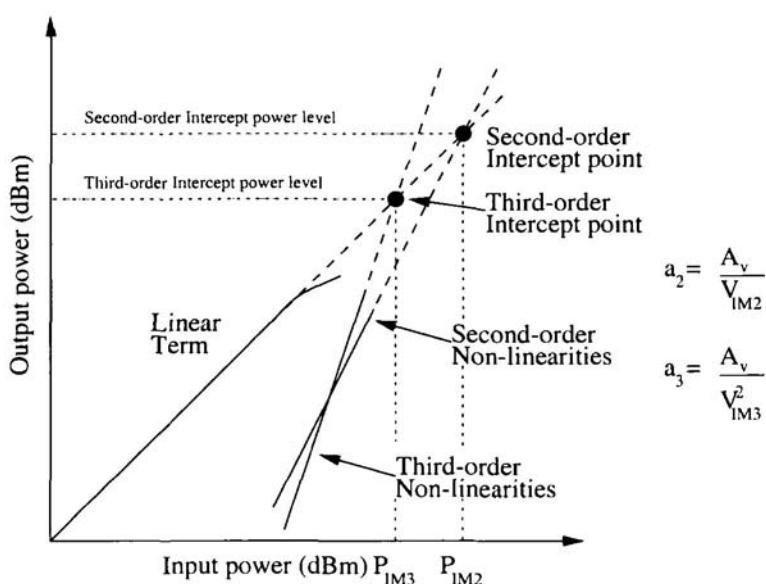


Figure 6.7. Intercept points for a third-order non-linear amplifier.

Assume that the small signal voltage gain of the amplifier, i.e.  $a_1$  of equation 6.13, is  $A_v$ .  $a_2$  and  $a_3$  are related to the intercept input voltages according to the following equations:

$$a_2 = \frac{A_v}{V_{IM2}} \quad (6.22)$$

$$a_3 = \frac{A_v}{V_{IM3}^2} \quad (6.23)$$

It should be noted that the intercept input voltage levels are outside the normal range of operation of the amplifier. In fact, the intercept points are derived analytically, by extrapolating the linear, second-order non-linear and third-order non-linear plots, as shown in Figure 6.7.

The third-order intercept is important for analyzing the in-band intermodulation distortion components ( $2f_1 - f_2$  and  $2f_2 - f_1$ ). While, for broadband systems, the second-order intercept point is used for determining the spur-free dynamic range.

### 6.3.1 Second-order intercept point of cascaded amplifiers

Consider two cascaded stages as shown in Figure 6.8. The first stage has a small signal voltage gain  $[A_v]_1$ , and a second-order intercept input voltage  $[V_{IM2}]_1$ . Then according to equations 6.13 and 6.22, the non-linear voltage transfer characteristics of the first stage is given by:

$$v_{out1} = [A_v]_1 v_{in} + \frac{[A_v]_1}{[V_{IM2}]_1} v_{in}^2 \quad (6.24)$$

Similarly, the voltage transfer characteristics of the second stage is given by:

$$v_{out} = [A_v]_2 v_{out1} + \frac{[A_v]_2}{[V_{IM2}]_2} v_{out1}^2 \quad (6.25)$$

Combining equations 6.24 and 6.25 together we get:

$$\begin{aligned} v_{out} &= [A_v]_1 [A_v]_2 v_{in} + \left\{ \frac{[A_v]_1 [A_v]_2}{[V_{IM2}]_1} + \frac{[A_v]_1^2 [A_v]_2}{[V_{IM2}]_2} \right\} v_{in}^2 \\ &\quad + \text{Higher order terms} \end{aligned} \quad (6.26)$$

Therefore, using equation 6.22, the overall second-order intercept input voltage of a two-stage receiver is given by:

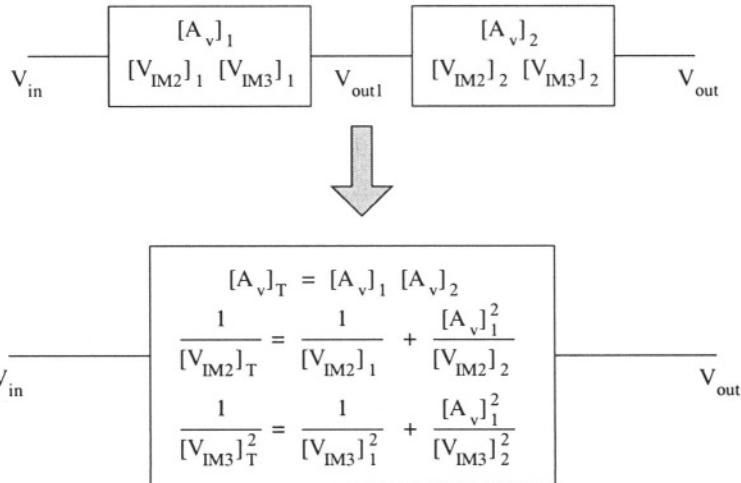


Figure 6.8. Intercept points of a two-stage non-linear receiver.

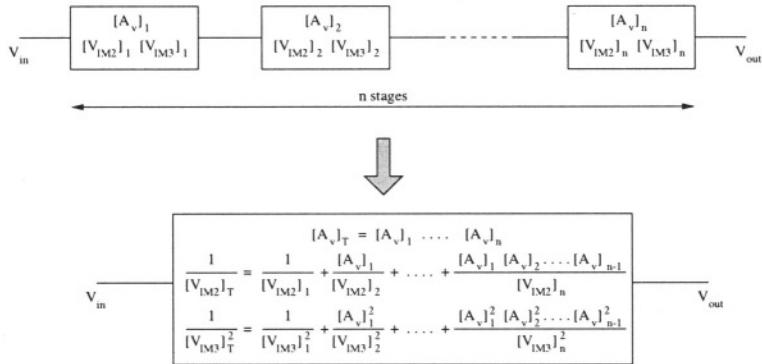


Figure 6.9. Intercept points of a  $n$ -stage non-linear receiver.

$$\frac{1}{[V_{IM2}]_T} = \frac{1}{[V_{IM2}]_1} + \frac{[A_v]_1}{[V_{IM2}]_2} \quad (6.27)$$

Expanding this to an  $n$ -stage receiver as that shown in Figure 6.9, we get:

$$\frac{1}{[V_{IM2}]_T} = \frac{1}{[V_{IM2}]_1} + \frac{[A_v]_1}{[V_{IM2}]_2} + \dots + \frac{[A_v]_1 [A_v]_2 \dots [A_v]_{n-1}}{[V_{IM2}]_n} \quad (6.28)$$

### 6.3.2 Third-order intercept point of cascaded amplifiers

Consider two cascaded stages as shown in Figure 6.8. The first stage has a small signal voltage gain  $[A_v]_1$ , and a third-order intercept input voltage

$[V_{IM_3}]_1$ . Then according to equations 6.13 and 6.23, the non-linear voltage transfer characteristics of the first stage is given by:

$$v_{out1} = [A_v]_1 v_{in} + \frac{[A_v]_1}{[V_{IM_3}]_1^2} v_{in}^3 \quad (6.29)$$

Similarly, the voltage transfer characteristics of the second stage is given by:

$$v_{out} = [A_v]_2 v_{out1} + \frac{[A_v]_2}{[V_{IM_3}]_2^2} v_{out1}^3 \quad (6.30)$$

Combining equations 6.29 and 6.30 together we get:

$$\begin{aligned} v_{out} &= [A_v]_1 [A_v]_2 v_{in} + \left\{ \frac{[A_v]_1 [A_v]_2}{[V_{IM_3}]_1^2} + \frac{[A_v]_1^3 [A_v]_2}{[V_{IM_3}]_2^2} \right\} v_{in}^3 \\ &\quad + \text{Higher order terms} \end{aligned} \quad (6.31)$$

Therefore, using equation 6.23, the overall third-order intercept input voltage of a two-stage receiver is given by:

$$\left( \frac{1}{[V_{IM_3}]_T} \right)^2 = \left( \frac{1}{[V_{IM_3}]_1} \right)^2 + \left( \frac{[A_v]_1}{[V_{IM_3}]_2} \right)^2 \quad (6.32)$$

Expanding this to an  $n$ -stage receiver as that shown in Figure 6.9, we get:

$$\begin{aligned} \left( \frac{1}{[V_{IM_3}]_T} \right)^2 &= \left( \frac{1}{[V_{IM_2}]_1} \right)^2 + \left( \frac{[A_v]_1}{[V_{IM_3}]_2} \right)^2 + \\ &\quad \dots + \left( \frac{[A_v]_1 [A_v]_2 \cdots [A_v]_{n-1}}{[V_{IM_3}]_n} \right)^2 \end{aligned} \quad (6.33)$$

## 6.4 THE SUPERHETERODYNE RECEIVER

In a superheterodyne receiver, the most commonly used receiver topology, the received signal is down-converted to a lower intermediate frequency (IF) by a mixer. Hence, this receiver is also known as the IF receiver. The IF frequency is a fixed frequency which makes filtering and detection easier than in the RF (radio frequency) band, where the center frequency depends on the channel being selected.

### 6.4.1 Single-stage superheterodyne receiver

Figure 6.10 shows a single stage IF receiver, where the received RF signal is down-converted to a lower IF frequency by multiplying it with a sinusoidal waveform. The RF signal, at the input to the receiver, has a center frequency  $f_{RF}$ , and is represented as:

$$v_{RF}(t) = m(t) \cos(2\pi f_{RF} t) \quad (6.34)$$

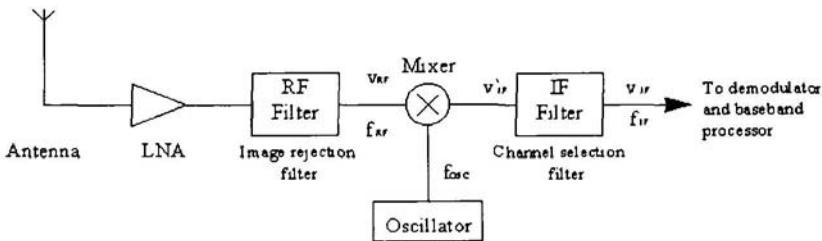


Figure 6.10. Block diagram of a single-stage superheterodyne receiver.

Where,  $m(t)$  is the modulating signal, that contains the information conveyed across the channel. The RF signal is multiplied by the sinusoidal waveform,  $2 \cos(2\pi f_{osc} t)$ , which has a center frequency  $f_{osc}$ , such that the generated IF signal can be expressed as:

$$\begin{aligned} v'_{IF}(t) &= 2 \cos(2\pi f_{osc} t) \times m(t) \sin(2\pi f_{RF} t) \\ &= m(t) \{ \cos[2\pi(f_{osc} - f_{RF})t] + \cos[2\pi(f_{osc} + f_{RF})t] \} \end{aligned} \quad (6.35)$$

The second term of equation 6.35 is filtered out by the IF filter. Hence, the output of the IF filter is:

$$v_{IF} = m(t) \cos(2\pi f_{IF} t) \quad (6.36)$$

Where,  $f_{IF}$  is the intermediate frequency, which is given by (for a high-side oscillator):

$$f_{IF} = f_{osc} - f_{RF} \quad (6.37)$$

Each desired RF frequency has an image frequency, also known as a mirror frequency. This image frequency, which is denoted by  $f_{im}$ , when mixed with the oscillator frequency  $f_{osc}$ , produces a component at the intermediate

frequency  $f_{IF}$ , as shown in Figure 6.11. While, for a high-side oscillator,  $f_{RF}$  is smaller than  $f_{osc}$ ,  $f_{im}$  is larger than  $f_{osc}$ , such that:

$$f_{im} = f_{osc} + f_{IF} \quad (6.38)$$

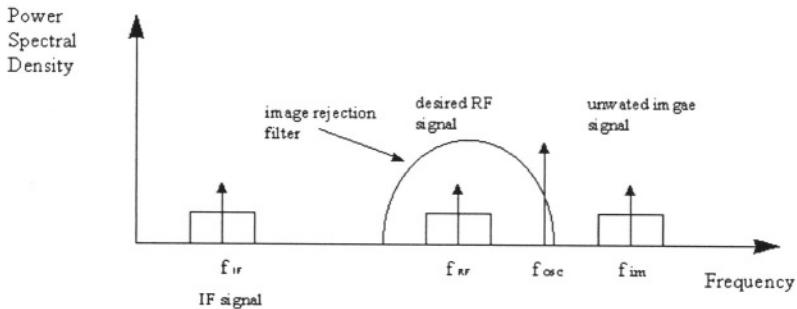


Figure 6.11. Image frequency in a superheterodyne receiver.

To prevent the image frequency from passing through, an image rejection filter is used before the mixer as shown in Figure 6.10. This filter, which is centered around the RF frequency, should have a high enough quality factor  $Q$  ( $Q$  is the ratio between the center frequency to the bandwidth of the filter) to suppress the image frequency.  $f_{IF}$  is usually one order of magnitude lower than  $f_{RF}$  (i.e.  $f_{RF}/f_{IF} = 10$ ). Hence, a filter having a quality factor  $Q$  of 50 to 100 is adequate for image rejection [36]. Such a filter is usually implemented as an off-chip filter. Another degree of complexity in the design of the image rejection filter is the fact that the center frequency of the filter might not be fixed.

The IF signal is further filtered by a low-bandwidth fixed-center frequency filter to select the desired channel. This filter too, is usually implemented as an off-chip filter. It should be noted that the choice of the intermediate frequency is a matter of compromise. On one hand, it is desirable to make  $f_{IF}$  large to have an image rejection filter with a large bandwidth and hence a low quality factor. While on the other hand it is desirable to make  $f_{IF}$  low to lower the center frequency of the channel selection filter and hence lower its quality factor. One way around this paradox is to use a multi-stage superheterodyne receiver, as explained in the following section.

#### 6.4.2 Multi-stage superheterodyne receiver

A multi-stage superheterodyne receiver alleviates the image frequency problem by doing down-conversion on a few stages, as shown in the two stage

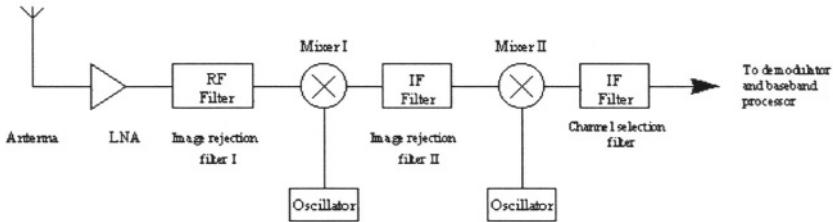


Figure 6.12. A two-stage superheterodyne receiver.

superheterodyne receiver of Figure 6.12. Each stage typically, does down-conversion by one decade, i.e. the center frequency at the output of the mixer is one-tenth that at its input.

A multi-stage superheterodyne receiver allows the use of low quality factor filters for image frequency rejection and channel selection. However, more filters are required as shown in Figure 6.12, where two filters are need for image frequency rejection, one before each mixer, in addition to the channel selection filter after the second stage. All these filters are usually implemented off-chip using discrete components.

### 6.4.3 Mixing Techniques

The simplest way to implement the down-conversion mixer is to multiply the incoming RF signal, which has a frequency  $f_{RF}$  by a sinusoidal signal generated by the local oscillator having a frequency  $f_{osc}$ . This mixer is shown in Figure 6.10. The output of the multiplier is a double sideband signal, having two frequency components at  $f_{osc} - f_{RF}$  and  $f_{osc} + f_{RF}$ . A low-pass filter is needed to filter out the upper side-band. Thus, the output of the low-pass filter is a sinusoidal signal having a frequency  $f_{osc} - f_{RF} = f_{IF}$ .

It is possible to construct the mixer in such a way that it suppresses the upper (or lower) sideband from the output, thus eliminating the need for a low-pass filter after the mixer. Figure 6.13 shows a block diagram of a mixer that generates only one sideband at its output. This mixer has two multipliers, the first one multiplies the input RF signal by the local oscillator signal, while the second one multiplies a  $90^\circ$  phase shifted version of the input signal by a  $90^\circ$  phase shifted local oscillator signal. The output of the two multipliers is given by:

$$\begin{aligned}
 v_{out1}(t) &= m(t) \cos(2\pi f_{RF} t) \cos(2\pi f_{osc} t) \\
 &= \frac{1}{2} m(t) \cos(2\pi [f_{osc} - f_{RF}] t)
 \end{aligned}$$

$$+ \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}} + f_{\text{RF}}]t) \quad (6.39)$$

$$\begin{aligned} v_{\text{out}2}(t) &= m(t) \sin(2\pi f_{\text{RF}} t) \sin(2\pi f_{\text{osc}} t) \\ &= \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}} - f_{\text{RF}}]t) \\ &\quad - \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}} + f_{\text{RF}}]t) \end{aligned} \quad (6.40)$$

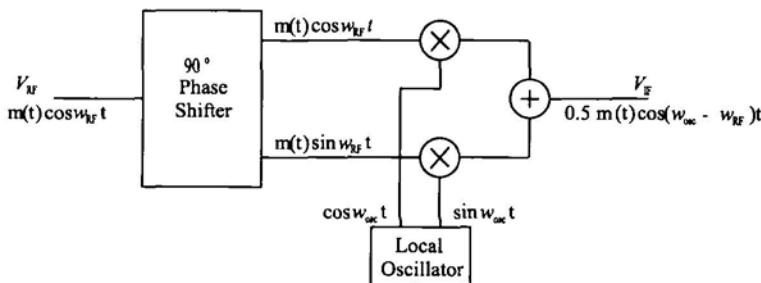


Figure 6.13. Single side-band mixer.

The output of each multiplier has two sidebands at frequencies  $f_{\text{osc}} - f_{\text{RF}}$  and  $f_{\text{osc}} + f_{\text{RF}}$ . The lower side-bands of each multiplier are in-phase, while the upper side-bands are out-of-phase. When we add the outputs of the two multipliers, the upper side-band cancels out, leaving only the lower side-band at the output of the mixer having a frequency  $f_{\text{osc}} - f_{\text{RF}}$ .

Figure 6.14 shows an alternative single side-band mixer that avoids the use of a phase shifting block, instead it uses two multipliers and two low-pass filters. The output signals of the low-pass filters have a  $90^\circ$  phase shift between them. The output of the two multipliers of the first stage are given by:

$$\begin{aligned} v_{\text{out}1} &= m(t) \cos(2\pi f_{\text{RF}} t) \cos(2\pi f_{\text{osc}1} t) \\ &= \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}1} + f_{\text{RF}}]t) \\ &\quad + \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}1} - f_{\text{RF}}]t) \end{aligned} \quad (6.41)$$

$$v_{\text{out}2} = m(t) \cos(2\pi f_{\text{RF}} t) \sin(2\pi f_{\text{osc}1} t)$$

$$\begin{aligned}
 &= \frac{1}{2}m(t) \sin(2\pi[f_{\text{osc}1} + f_{\text{RF}}]t) \\
 &\quad + \frac{1}{2}m(t) \sin(2\pi[f_{\text{osc}1} - f_{\text{RF}}]t)
 \end{aligned} \tag{6.42}$$

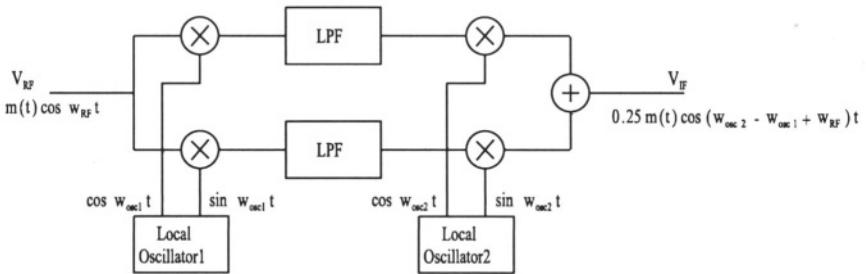


Figure 6.14. Alternative single side-band mixer.

The low-pass filters remove the upper sideband leaving only the lower sideband of each signal as given by the following equations:

$$v_{out3} = \frac{1}{2}m(t) \cos(2\pi[f_{\text{osc}1} - f_{\text{RF}}]t) \tag{6.43}$$

$$v_{out4} = \frac{1}{2}m(t) \sin(2\pi[f_{\text{osc}1} - f_{\text{RF}}]t) \tag{6.44}$$

These two signals are then multiplied by two quadrature signals generated by a second local oscillator. The output of the two second-stage multipliers are given by:

$$\begin{aligned}
 v_{out5} &= \frac{1}{4}m(t) \cos(2\pi[f_{\text{osc}2} - f_{\text{osc}1} + f_{\text{RF}}]t) \\
 &\quad + \frac{1}{4}m(t) \cos(2\pi[f_{\text{osc}2} + f_{\text{osc}1} - f_{\text{RF}}]t)
 \end{aligned} \tag{6.45}$$

$$\begin{aligned}
 v_{out6} &= \frac{1}{4}m(t) \cos(2\pi[f_{\text{osc}2} - f_{\text{osc}1} + f_{\text{RF}}]t) \\
 &\quad - \frac{1}{4}m(t) \cos(2\pi[f_{\text{osc}2} + f_{\text{osc}1} - f_{\text{RF}}]t)
 \end{aligned} \tag{6.46}$$

Adding these two signals together leaves a single frequency component at the output of this mixer, whose frequency is given by,  $f_{\text{osc}2} - f_{\text{osc}1} + f_{\text{RF}}$ .

## 6.5 THE HOMODYNE RECEIVER

The homodyne receiver, also known as the zero-IF or direct conversion receiver, down-converts the received signal directly from RF to baseband as shown in Figure 6.15. The band-pass RF signal is a quadrature modulated signal having both in-phase and quadrature-phase components, to be able to recover both components quadrature down-conversion is used. Assume that the received signal  $r(t)$  is given by:

$$r(t) = m_i(t) \cos(w_c t) + m_q(t) \sin(w_c t) \quad (6.47)$$

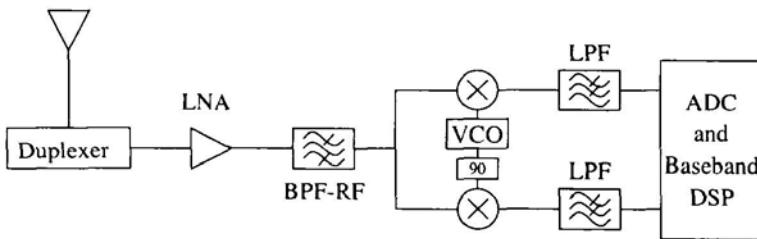


Figure 6.15. The homodyne (zero-IF) receiver.

Quadrature down-conversion involves multiplying the received signal  $r(t)$  by  $\cos(w_c t)$  and  $\sin(w_c t)$ .  $b_i(t)$  and  $b_q(t)$  are the outputs of each mixer respectively.

$$\begin{aligned} b_i(t) &= r(t) \cos(w_c t) \\ &= \frac{1}{2}m_i(t)[1 + \cos(2w_c t)] + \frac{1}{2}m_q(t)\sin(2w_c t) \end{aligned} \quad (6.48)$$

$$\begin{aligned} b_q(t) &= r(t) \sin(w_c t) \\ &= \frac{1}{2}m_q(t)[1 - \cos(2w_c t)] + \frac{1}{2}m_i(t)\sin(2w_c t) \end{aligned} \quad (6.49)$$

Using a low-pass filter after the mixer eliminates the high frequency components centered around  $2f_c$ , and leaves the baseband components.

There are several advantages for using a homodyne receiver:

1. The image frequency problem is eliminated, hence there is no need to use high quality filters in the front end of the receiver.
2. The filters used for channel selection are low-pass filters, which can easily be integrated, instead of using high quality factor band-pass filters, as in the case of the superheterodyne receiver.

It should be noted that in the homodyne receiver a front-end filter is still needed to improve the dynamic range performance of the receiver, by eliminating out of band interference. This filter is also needed to avoid the oscillator's harmonics from mixing unwanted RF signals into baseband. Without a front-end band-pass filter, the RF signals present around  $f_c$ ,  $2f_c$ ,  $3f_c$ , etc. are all aliased into baseband at the output of the low-pass filter of Figure 6.15. This is shown in Figure 6.16. However, using a low-pass filter before the mixer, having a cutoff frequency between  $f_c$  and  $2f_c$ , also prevents the frequency components at the harmonics of  $f_c$  from aliasing into baseband.

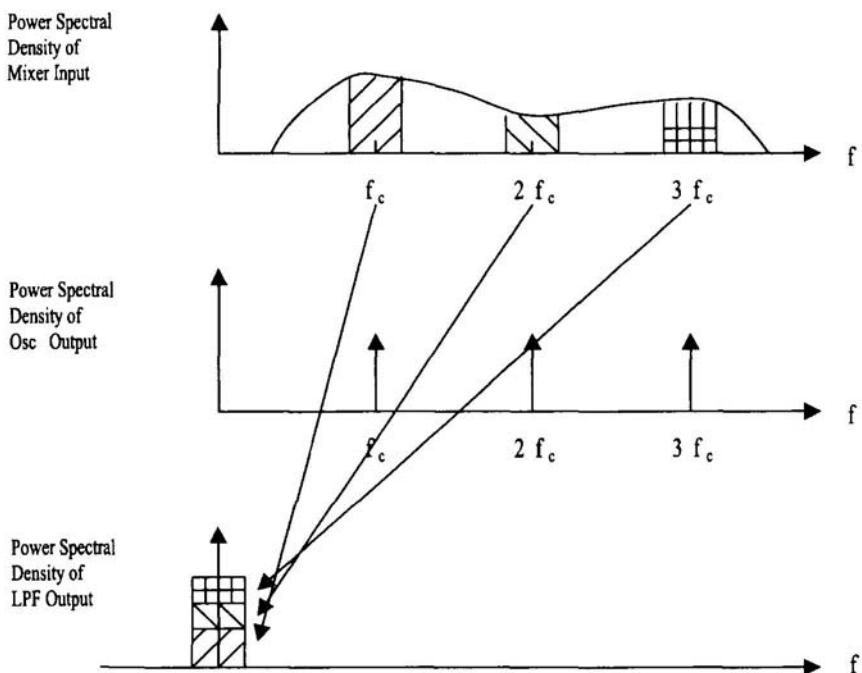


Figure 6.16. Aliasing in a homodyne receiver.

## 6.6 SOFTWARE RADIO

In software radio [37], [38], the received/transmitted radio signal is digitized as electrically close to the antenna as possible, converting it into a digital signal at a very high frequency. The signal processing is done digitally after that, using general purpose programmable hardware. Performing the radio, IF and baseband functions in programmable digital hardware increases the flexibility of the transceiver. Although software radios use digital techniques,

digital radios are generally not software radios. The key difference is the total programmability of software radios, including programmable RF bands, multiple access modes, and modulation schemes.

Software radio was conceived in the 70's. However, technology limitations prevented it from being implemented. The first operational digital high frequency communications system was built in 1980 [39]. It was used by the military. That system occupied many racks, it dissipated a large amount of power, and it only had a bandwidth, for simultaneous coverage, of 750 KHz, with a dynamic range of 60 dB.

Today the US military is in phase II of developing its software radio Speakeasy [38]. Speakeasy is a programmable multi-band multi-mode radio (MBMMR) that operates in the HF to the UHF bands, from 2 MHz to 2 GHz. Speakeasy emulates 15 existing military radios. It supports 9 modulation schemes, and 4 digital audio coding algorithms. It also supports multiple internetworking protocols, multiple interfaces, multiple forward error correction codes and multiple information security (INFOSEC) algorithms.

For civil applications, software radio is used in cutting-edge base stations. The design of portable terminals is a compromise between low-power and high-performance, this involves a tradeoff between analog ICs, low-power ASICs, DSP cores and embedded microprocessors [37]. However, as low-power techniques and design methodologies emerge, digital signal processing will gradually replace analog signal processing in the wireless portable terminal.

There are numerous advantages to increasing the portion of the radio that is implemented digitally. These include relaxing the analog components requirements. Digital implementations tend to be compact and inexpensive for large volume production. One of the most important advantages is the ability to program digital structures to meet the communication needs of different networks using a single hardware platform.

The access, modulation and coding schemes used in a software radio are programmable, making it possible to reprogram the transceiver if any of these schemes change. Channel selection, propagation channel characterization, antenna steering and power level adjustment are all done under software control [37]. In the transmit mode, the software radio characterizes the available channels, steers the transmit beam in the right direction, selects the appropriate power level and then transmits the signal. In the receive mode, the software radio analyzes the received spectrum, in frequency, time and space. It identifies the interferers and nulls them. It estimates the multi-path propagation channel model and adaptively equalizes the received signal. The signal is then demodulated and decoded.

Software radio is characterized by its modular, open architecture allowing constant upgrades as the technology advances. The software radio architecture, as shown in Figure 6.17, consists of three subsystems. The real-time channel

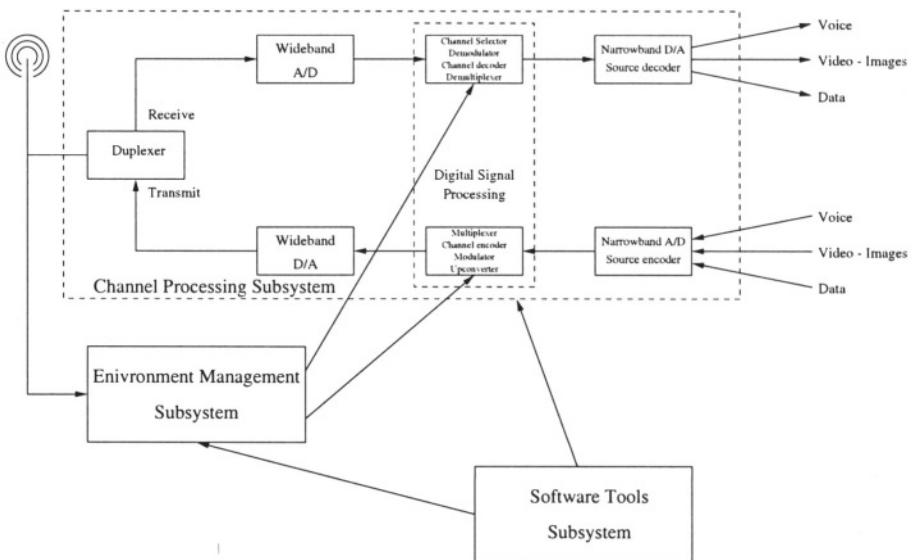


Figure 6.17. The software radio architecture.

processing subsystem is where all the radio functions are performed. This subsystem must have isochronous performance, which means that the input samples must be processed during a limited time duration. The environment management subsystem constantly characterizes the radio environment. This information is used by the channel processing subsystem for better transmission and reception. The environment management subsystem has near real-time operation. The software tools subsystem provides incremental service enhancements. This subsystem allows, defining, prototyping, testing and delivering these service enhancements.

There are two bottlenecks that challenge the implementation of software radio. The first is the need for high-speed and high-resolution analog-to-digital converters. To implement software radio the entire spectrum of a particular standard should be digitized. This is a 25 MHz band for GSM, IS-136 and IS-95, as given in Table 1.1. To digitize a 25 MHz band-pass signal, band-pass sampling is used [40]. To satisfy the Nyquist sampling criteria, the sampling frequency should be at least twice the bandwidth. Assuming the sampling frequency is 2.5 times the bandwidth, then a sampling frequency of 62.5 MSa/s is required. To meet the requirements of the different wireless standards the analog-to-digital converter is required to have over 20 bits resolution, this is needed to achieve the desired dynamic range and selectivity in the receiver.

The second bottleneck that challenges the implementation of software radio is the high DSP processing power required. Typically Software Radio requires

up to 10 GFLOPS/s [41]. Such high processing power is beyond the capabilities of today's DSPs.

## Chapter 7

# ANALOG TO DIGITAL CONVERSION

### 7.1 INTRODUCTION

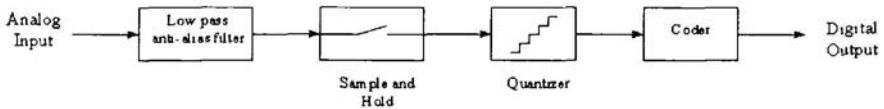
The advancement of VLSI technology led to the proliferation of digital integrated circuits and systems which perform intricate signal processing in the digital domain.

Compared with analog signal processors, digital signal processors (DSPs) have numerous advantages. In digital systems, the signal is quantized into discrete levels, and a finite number of digital code-words are transmitted, most of the noise and interference added to the digital signal during processing or transmission can be removed. However, in analog systems any noise added to the signal is indistinguishable from it and hence can't be removed. Therefore, analog signal processing requires accurate components with precise tolerance. However, digital signal processing can tolerate less precise components making digital signal processors less susceptible to temperature, aging and manufacturing tolerances. Furthermore, digital systems allow more intricate signal processing and offer more extensive programmability than analog systems.

However, all naturally occurring signals that are encountered in the real world are analog signals. This necessitates the transformation of such signals from the analog domain to the digital domain to make use of the powerful computational processing power of the digital signal processors. The digital signal then has to be transformed back to the analog domain. This transformation is done by using analog-to-digital and digital-to-analog converters.

Figure 7.1 shows the block diagram of an analog-to-digital converter (ADC). The low-pass filter, which is known as the anti-alias filter, band-limits the analog signal so as to prevent aliasing from occurring in the sampler. The sampler discretizes the signal in the time domain. This is then followed by the quantizer, which is a many-to-one transformer that maps a range of the

continuous signal into a discrete level. The quantizer performs approximation to the analog signal by approximating it to one of a finite number of discrete levels. After being quantized, the coder maps each quantized level into a binary code-word.



*Figure 7.1.* Block diagram of an analog-to-digital converter (ADC).

In the digital-to-analog converter (DAC), the reverse operations to those of the analog-to-digital converter occur as shown in Figure 7.2. The decoder transforms the binary code into a quantized signal level. Because the quantizer is a many to one transformer, i.e. it maps a range of the continuous signal into a discrete level, hence, it has no inverse equivalent in the digital-to-analog converter. Thus, any quantization noise added to the signal is stuck to it and can't be removed by the digital-to-analog converter. Finally, a low-pass filter converts the time-discrete (sampled) signal into a continuous analog signal.



*Figure 7.2.* Block diagram of a digital-to-analog converter (DAC).

Figure 7.3 shows a block diagram of a digital communications system. The signal to be transmitted is an analog signal, usually speech. This signal is digitized by an analog-to-digital converter. The digital signal is processed by a digital signal processor, that performs functions such as source encoding, channel encoding, time division multiplexing - for a TDMA system, and code spreading for a CDMA system.

After digital processing the signal is converted back into the analog domain by a digital-to-analog converter. The converted analog signal can be at baseband or at IF (Intermediate Frequency). Further analog signal processing is performed on the analog signal by an analog signal processor. This processing includes, frequency up conversion to the RF band, filtering, and power amplification.

The analog signal received at the receiver is processed by an analog signal processor that does low noise amplification, filtering, and frequency down

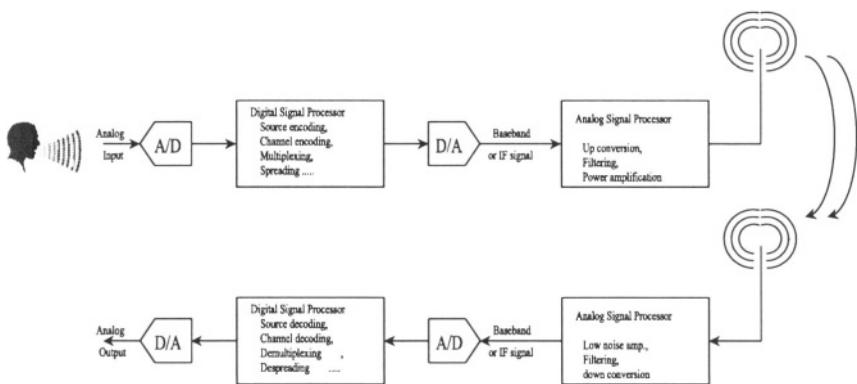


Figure 7.3. Block diagram of a digital communications system.

conversion to IF or to baseband. The signal is then digitized and processed by a digital signal processor that despreading, demultiplexes and decodes the signal. Eventually, the digital signal is transformed back into the analog domain to get the received replica of the transmitted signal.

An important decision in the design of a digital communications system is the partitioning of the signal processing between the analog and digital domains. Despite the powerful capabilities of digital processors, yet some signal processing functions remain to be in the exclusive territory of analog signal processors because of complexity and power dissipation considerations. However, as DSPs become more powerful and as their power dissipation continues to drop with voltage and technology scaling. The digital creep into the analog domain is expected to continue. This necessitates the use of high-speed high-resolution signal conversion devices (ADCs and DACs).

In this chapter we presented different types of analog-to-digital converters. The performance metrics of ADCs are discussed in section 7.2. The first stage of analog-to-digital conversion after low-pass filtering is sampling, which converts a time-continuous analog signal into a discrete signal in the time-domain. Sampling is presented in section 7.3. Band-pass signals can be sampled at a rate lower than the Nyquist rate for band-limited signals. Band-pass sampling is the topic of section 7.4. At the heart of any analog-to-digital converter is the quantizer, which is responsible for converting analog samples into discrete (digital) samples. Quantization is presented in section 7.5. In section 7.6, we discuss different types of analog-to-digital converters such the flash, folding, interpolative, successive approximation and pipeline ADCs. Finally, in section 7.7 we present the Sigma-Delta analog-to-digital converter,

which is a two-stage oversampling ADC. The first stage converts the analog signal into an oversampled low-resolution digital signal, the second stage converts the oversampled low-resolution digital signal into a Nyquist-rate high-resolution digital signal.

## 7.2 PERFORMANCE METRICS OF ANALOG-TO-DIGITAL CONVERTERS

The following parameters define the performance of an analog-to-digital converter:

**Sampling Rate.** This is the number of times the analog signal is sampled per second.

**Signal-to-Noise Ratio (SNR).** This is the ratio between the signal power to the quantization noise. This ratio depends on the number of bits at the output of the analog-to-digital converter.

**Effective Number of Bits (ENOB).** The effective number of bits is related to the signal-to-noise ratio by the following equation:

$$\text{ENOB} = \frac{\text{SNR(dB)} - 1.76(\text{dB})}{6(\text{dB})} \text{ (bits)} \quad (7.1)$$

**Dynamic Range.** This is the ratio between the maximum input power and the minimum input power. The maximum input is measured at the onset of overload. While the minimum input power is measured when the output signal-to-noise is equal to 0 dB.

**DC Offset.** This is the point at which the straight line passing through the transfer characteristics of the analog-to-digital converter, as shown in Figure 7.4, cuts the vertical axis. Ideally, this value should be zero.

**Gain.** This is the slope of the straight line passing through the transfer characteristics of the analog-to-digital converter. Ideally, this value should be one.

**Differential Non-linearity (DNL).** This is the maximum deviation in the step size from the ideal step size value.

**Integral Non-linearity (INL).** This is the maximum deviation of the transfer characteristics of the analog-to-digital converter from a straight line passing through the end points of the transfer characteristics.

## 7.3 SAMPLING

Sampling is the process of converting a continuous analog signal into a discrete-time signal or a sequence of numbers. Depending on the characteristics of the sampling circuit, sampling can be modeled differently, resulting in different frequency spectra for the sampled signal. In this section, we consider three sampling models as shown in Figure 7.5:

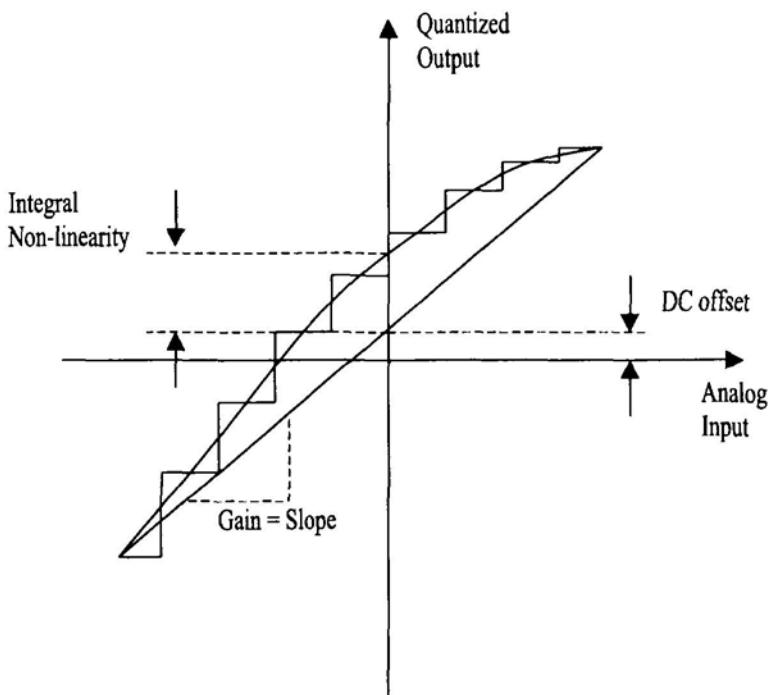


Figure 7.4. Non-linear transfer characteristic of an analog-to-digital converter.

- Ideal sampling.** This is shown in Figure 7.5.a, the sampled signal is the product of the analog signal and an impulse train. This type of sampling is also known as impulse sampling or instantaneous sampling.
- Pulse sampling.** This is shown in Figure 7.5.b, the sampled signal is the product of the analog signal and a pulse train.
- Flat-top sampling.** This is shown in Figure 7.5.c, in this case, the sampler captures the value of the analog signal at a particular time instance, and then holds this value constant for the duration of the pulse. Flat-top sampling is modeled mathematically as multiplication by an impulse train followed by a convolution with a pulse.

### 7.3.1 Ideal Sampling

Consider a band-limited low-pass signal  $m(t)$  as shown in Figure 7.6.a. Assume that  $M(f)$  is the frequency domain representation of that signal. Suppose that we want to sample this signal every  $T_s$  seconds, this is achieved

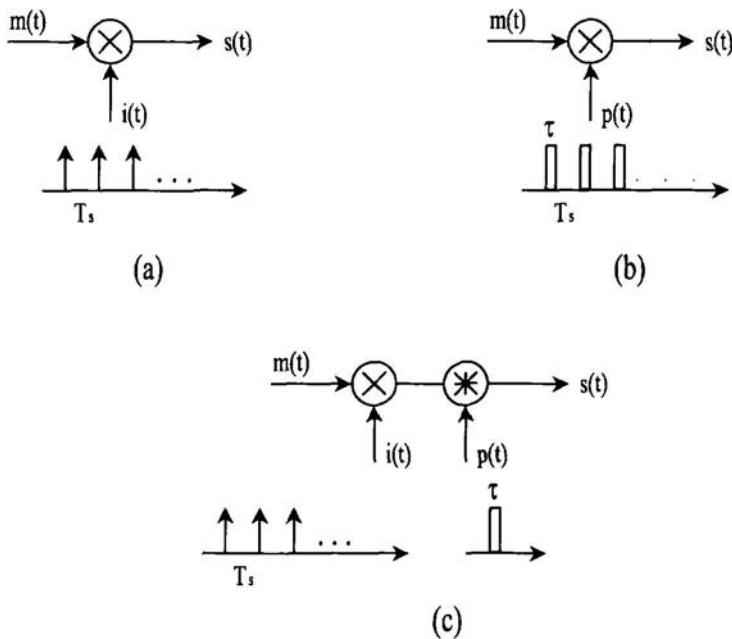


Figure 7.5. Sampling models: (a) Ideal sampling. (b) Pulse sampling. (c) Flat-top sampling.

by multiplying the continuous input signal  $m(t)$  by an infinite impulse train  $i(t)$  having a period  $T_s$ , as shown in Figure 7.6.  $i(t)$  is given by:

$$i(t) = \sum_{n=-\infty}^{\infty} \delta(t - nT_s) \quad (7.2)$$

The frequency domain transformation of  $i(t)$  is also an impulse train given by:

$$I(f) = \frac{1}{T_s} \sum_{m=-\infty}^{\infty} \delta\left(f - \frac{m}{T_s}\right) \quad (7.3)$$

Where  $\frac{1}{T_s}$  is the sampling frequency  $f_s$ .

The sampled signal  $s(t) = m(t) \times i(t)$  is given by:

$$s(t) = \sum_{n=-\infty}^{\infty} m(nT_s) \delta(t - nT_s) \quad (7.4)$$

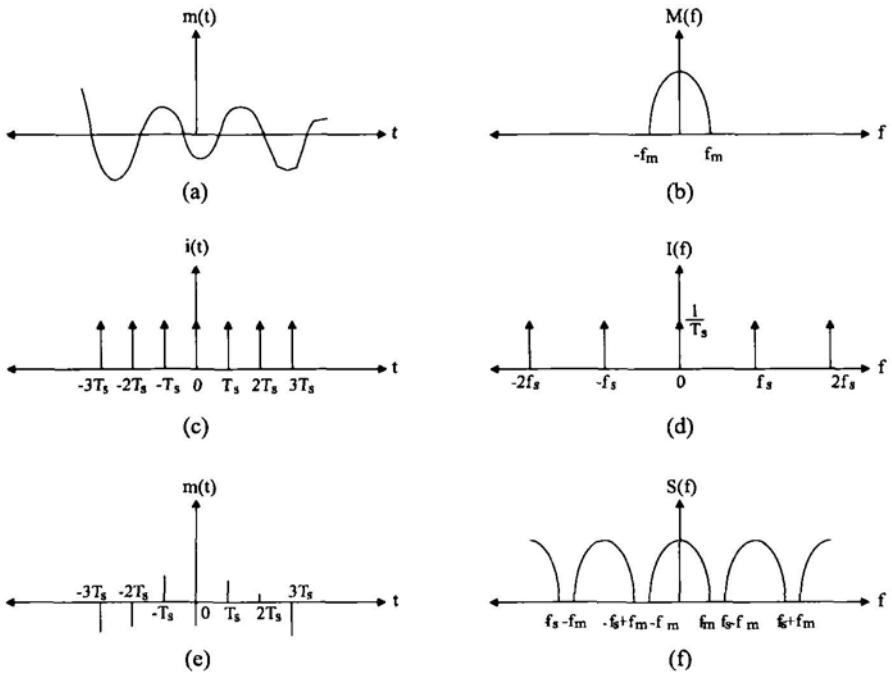


Figure 7.6. Ideal sampling: (a) Time-domain representation of the input signal. (b) Frequency response of the input signal. (c) Time-domain representation of the impulse train. (d) Frequency response of the impulse train. (e) Time-domain representation of the sampled signal. (f) Frequency response of the sampled signal.

Multiplication in the time domain corresponds to convolution in the frequency domain. Hence, the frequency domain transformation of the sampled signal is given by:

$$S(f) = \frac{1}{T_s} \sum_{m=-\infty}^{\infty} M\left(f - \frac{1}{T_s}\right) \quad (7.5)$$

Sampling produces images of the frequency spectrum of the band-limited signal ( $M(f)$ ) at multiples of the sampling frequency  $f_s$ , as shown in Figure 7.6.f. To recover the continuous band-limited signal  $m(t)$ , the sampled signal  $s(t)$  is passed through an ideal low-pass filter having a frequency response:

$$H(f) = \begin{cases} T_s & |f| \leq f_m \\ 0 & \text{otherwise} \end{cases} \quad (7.6)$$

Where  $f_m$  is the maximum frequency of the band-limited signal.

To perfectly reconstruct the continuous band-limited signal  $m(t)$ , from the sampled signal  $s(t)$ , the maximum frequency  $f_m$  of the band-limited signal must be less than or equal half the sampling frequency  $f_s$ , i.e.

$$2f_m \leq f_s \quad (7.7)$$

Inequality 7.7 is known as the Nyquist condition for perfect reconstruction of a band-limited signal. The minimum sampling frequency that satisfies inequality 7.7 is known as the Nyquist rate. If the Nyquist condition of 7.7 is not satisfied, the spectra of the images overlap causing aliasing, as shown in Figure 7.7. When this occurs the signal  $m(t)$  can't be recovered from the sampled signal.

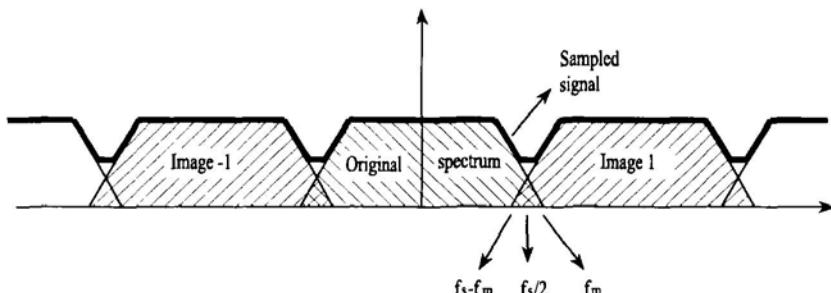


Figure 7.7. The spectrum of a sampled-signal sampled at less than the Nyquist rate.

In practice, a low-pass filters having an ideal transfer function as given by equation 7.6 is not realizable. A low-pass filter has a transition band from the pass-band to the stop-band. This necessitates leaving a guard band between the upper frequency of the signal band ( $f_m$ ), and the lower frequency of the adjacent image ( $f_s - f_m$ ). Thus, the guard band  $\Delta$  is given by:

$$\Delta = f_s - 2f_m \quad (7.8)$$

If the input time-continuous signal to the sampler is not band-limited, it is passed first through an anti-alias low-pass filter, this removes all frequency components above a certain maximum frequency  $f_m$ .  $f_m$  is chosen such that the energy of the input continuous signal above  $f_m$  is a small percentage of the total energy of the signal. The band-limited time-continuous signal can now be sampled at a sampling rate  $f_s$  not less than  $2f_m$ .

The sampling method shown in Figure 7.6 is known as ideal sampling. It is not practical to realize this type of sampling. A realizable sampling method replaces the impulses by finite duration pulses. This sampling model is discussed in the next section.

### 7.3.2 Pulse Sampling

In pulse sampling, the sampling signal is a pulse train as shown in Figure 7.5.b. The pulse train is defined mathematically as:

$$p(t) = \sum_{n=-\infty}^{\infty} p_{\tau}(t - nT_s) \quad (7.9)$$

Where,

$$p_{\tau}(t) = \begin{cases} 1 & |t| \leq \frac{\tau}{2} \\ 0 & \text{otherwise} \end{cases} \quad (7.10)$$

The frequency transform of the pulse train of equation 7.9 is a sequence of impulses given by:

$$P(f) = \sum_{m=-\infty}^{\infty} \frac{\sin(\pi m \tau / T_s)}{\pi m} \delta\left(f - \frac{m}{T_s}\right) \quad (7.11)$$

The frequency spectra of the sampled signal,  $s(t) = m(t) \times p(t)$  is the convolution of the frequency spectra of the input time-continuous signal and the pulse train:

$$\begin{aligned} S(f) &= M(f) * P(f) \\ &= \sum_{m=-\infty}^{\infty} \frac{\sin(\pi m \tau / T_s)}{\pi m} M\left(f - \frac{m}{T_s}\right) \end{aligned} \quad (7.12)$$

The frequency spectrum of the sampled signal is shown in Figure 7.8. Notice that, in this sampling model, the frequency spectrum of the time-continuous signal  $M(f)$  is replicated at multiples of the sampling frequency ( $f_s = 1/T_s$ ), this is similar to ideal sampling. However, unlike ideal sampling, each frequency image is scaled by a different factor  $\sin(\pi m \tau / T_s) / (\pi m)$ .

As long as the sampling frequency  $f_s$  is more than twice the maximum frequency  $f_m$  of the band-limited signal, it is possible to recover the band-limited signal from the sampled signal by passing the sampled signal through a low-pass filter with a cutoff frequency  $f_c$  given by:

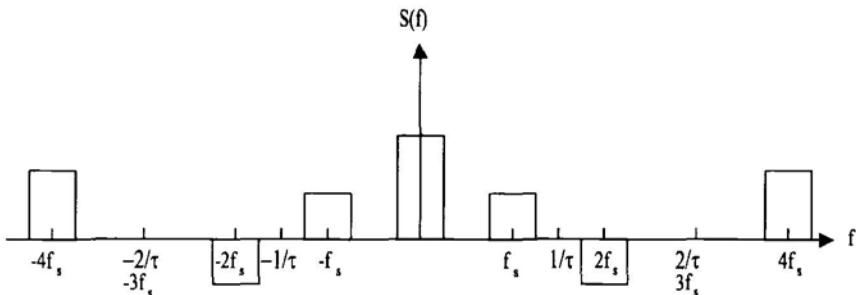


Figure 7.8. Frequency spectrum of a pulse sampled signal.

$$f_m \leq f_c \leq f_s - f_m \quad (7.13)$$

### 7.3.3 Flat-top sampling

In flat-top sampling, the sampler captures the value of the analog band-limited signal at a particular time instance, and then holds this value constant for the duration of the pulse ( $\tau$ ). Flat-top sampling is modeled mathematically as multiplication by an impulse train followed by convolution with a pulse of width  $\tau$ . Hence, the sampled signal ( $s(t)$ ) can be expressed in terms of the band-limited time-continuous analog signal ( $m(t)$ ) as:

$$s(t) = [m(t) \times i(t)] * p_\tau(t) \quad (7.14)$$

Where  $i(t)$  is the impulse train given by equation 7.2.

In the frequency domain, the sampled signal is expressed as:

$$S(f) = \sum_{n=-\infty}^{\infty} \frac{1}{T_s} M(f - mf_s) \cdot (\tau \text{sinc}(f\tau)) \quad (7.15)$$

Where,

$$\text{sinc}(x) = \frac{\sin(\pi x)}{\pi x}$$

The frequency spectrum of sampled signal is a distorted replica of the frequency spectrum of the band-limited time-continuous signal. The frequency spectrum of the band-limited time-continuous signal is repeated every  $f_s$ . The distortion happens because of the multiplication by the term ( $\tau \text{sinc}(f\tau)$ ) in the frequency-domain. Hence, to recover the band-limited time-continuous signal

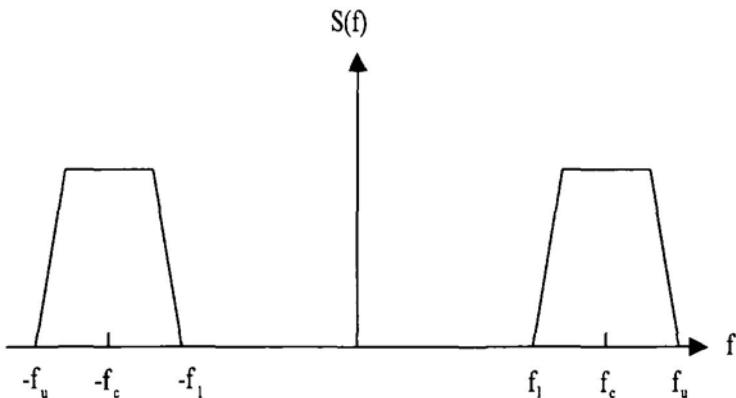


Figure 7.9. Frequency spectrum of a band-pass signal.

from the sampled signal, it isn't enough to use a low-pass filter, but an addition filter having a frequency response:

$$\frac{1}{\tau \text{sinc}(f\tau)}$$

is also required to remove the distortion introduced by flat-top sampling.

## 7.4 BAND-PASS SAMPLING

A band-pass signal is a signal whose frequency content is concentrated in a narrow-band that extends from  $f_l$  to  $f_u$  as shown in Figure 7.9. The band-pass signal has a center frequency  $f_c$  given by:

$$f_c = \frac{f_u + f_l}{2} \quad (7.16)$$

The bandwidth of the band-pass signal is 2 BW, where BW is given by:

$$\text{BW} = \frac{f_u - f_l}{2} \quad (7.17)$$

The sampling theory for band-limited signals, as given by equation 7.7, states that a sampling frequency of at least double the maximum frequency of the signal ( $f_s = 2f_u$ ) is needed to sample a band-limited signal. However, because of the band-pass nature of the band-limited signal, it is possible to sample it at a lower rate.

For every real band-pass signal  $m_b(t)$ , there exists an equivalent complex low-pass signal  $m_l(t)$ , having a bandwidth, BW, as given by equation 7.17. In the time domain, these signals are related by the following expression:

$$m_b(t) = \frac{1}{2}m_l(t)e^{j2\pi f_c t} + \frac{1}{2}m_l^*(t)e^{-j2\pi f_c t} \quad (7.18)$$

Where,

$*$  denotes the complex conjugate operation.

$f_c$  is the center frequency of the band-pass signal, which is given by equation 7.16.

Assume that  $M_b(f)$  and  $M_l(f)$  are the frequency transforms of  $m_b(t)$  and  $m_l(t)$  respectively. Hence, they are related by the following expression:

$$M_b(f) = \frac{1}{2}M_l(f - f_c) + \frac{1}{2}M_l(-f - f_c) \quad (7.19)$$

The equivalent low-pass signal can be sampled at a minimum rate of  $f_u - f_l$  for its real and imaginary components. Thus, the total real sample rate is  $2(f_u - f_l) = 2\text{BW}$ . If we were to sample the band-pass signal directly, the minimum sampling rate does not depend on the bandwidth, BW, but it depends on  $f_u$ , and  $f_l$  as well. Only under certain circumstance, can the minimum sampling frequency for the band-pass signal be 2 BW.

In general, to subsample a band-pass signal and ensure that spectrum overlap doesn't occur, the sampling frequency has to satisfy the following inequality [40]:

$$\frac{2f_u}{k} \leq f_s \leq \frac{2f_l}{k-1} \quad (7.20)$$

Where  $k$  is an integer satisfying the following inequality:

$$1 \leq k \leq \frac{f_c}{\text{BW}} + 0.5 \quad (7.21)$$

According to inequality 7.20, the sampling frequency depends on both, the bandwidth and the band position of the band-pass signal. Figure 7.10 [42], shows the valid sampling frequency as a function of the bandwidth BW and the center frequency  $f_c$  of the band-pass signal. Generally, the sampling frequency is given by [43]:

$$f_s = \frac{4f_c}{2k-1} \quad (7.22)$$

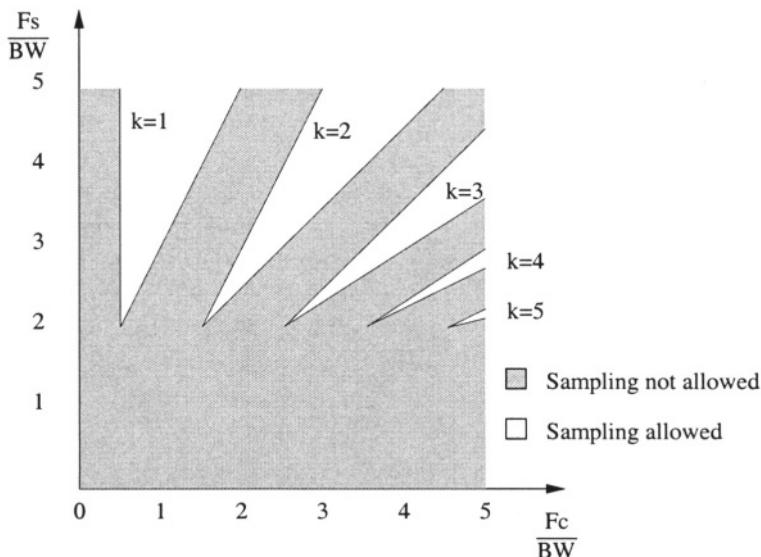


Figure 7.10. Allowable band-pass sampling rates.

## 7.5 QUANTIZATION

Quantization is the process of approximating a continuous analog signal to discrete levels, i.e. quantization is a many-to-one mapping. If these levels are uniformly spaced, the quantizer is a uniform quantizer. However, if these levels are unevenly spaced, the quantizer is a non-uniform quantizer.

Assume that the separation between any two consecutive quantization levels is the step-size  $\Delta$ . In a uniform quantizer  $\Delta$  is constant. Assume that,  $Q$  is the total number of quantization levels. In an analog-to-digital converter each quantization level is mapped into a binary power code. Hence,  $Q$  is a power of 2:

$$Q = 2^n \quad (7.23)$$

Where  $n$  is the number of bits at the output of the analog-to-digital converter. Figures 7.11 and 7.12 show two examples for the transfer function of the quantizer. Each transfer function consists of  $Q$  treads, each separated by  $\Delta$  and having a width  $\Delta$ .

The quantizer of Figure 7.11 is known as a mid-tread quantizer, this is because the origin is at the middle of a tread. The transfer function for this quantizer can be expressed by the following expression:

$$\text{If } (m - 1)\Delta < V_{in} < m\Delta \implies V_{out} = (m - \frac{1}{2})\Delta \quad (7.24)$$

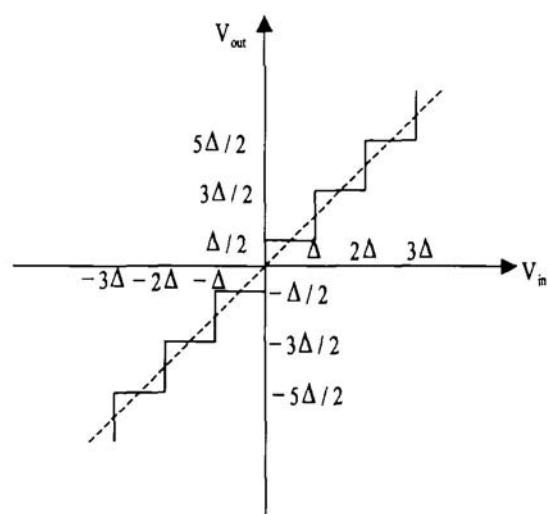


Figure 7.11. Staircase transfer function of a mid-tread quantizer.

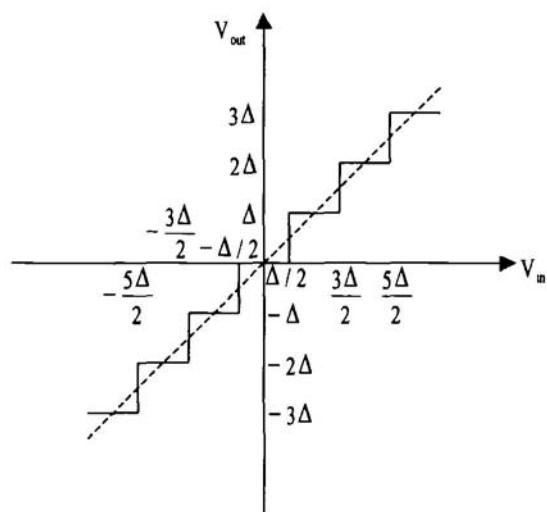


Figure 7.12. Staircase transfer function of a mid-rise quantizer.

Where,  $m = -\frac{Q}{2} + 1, -\frac{Q}{2} + 2, \dots, \frac{Q}{2} - 1, \frac{Q}{2}$ .

Notice that, for the mid-tread quantizer, the quantization levels are symmetric around the zero. However, there is no zero quantization level, i.e. the quantized signal is either positive or negative.

The quantizer of Figure 7.12 is known as a mid-rise quantizer, this is because the origin is at the middle of a rise. The transfer function for this quantizer can be expressed by the following expression:

$$\text{If } (m - \frac{1}{2})\Delta < V_{in} < (m + \frac{1}{2})\Delta \Rightarrow V_{out} = m\Delta \quad (7.25)$$

Where,  $m = -\frac{Q}{2}, -\frac{Q}{2} + 1, \dots, \frac{Q}{2} - 2, \frac{Q}{2} - 1$ .

Notice that, for the mid-rise quantizer, there exists a zero quantization level. However, in this case the quantization levels are not symmetric around the zero. There exists  $Q/2$  negative quantization levels, and only  $(Q/2 - 1)$  positive quantization levels. It is possible to make a the mid-rise quantizer symmetric by removing the most negative quantization level. However, in this case, the total number of quantization levels will not be a power of two.

Since quantization is an approximation process, it therefore introduces error into the quantized signal, which is known as the quantization error or quantization noise. Assume that  $v_{in}$  is the continuous analog input to the quantizer, and that  $v_{out}$  is the discrete output from the quantizer. The quantization error is defined as:

$$q = v_{in} - v_{out} \quad (7.26)$$

Figure 7.13 shows the quantization error for an 8-level mid-tread quantizer, versus the input signal. We can identify two regions in this Figure. The first region is where the input signal is large in magnitude ( $|v_{in}| > \frac{Q}{2}\Delta$ ). Where  $Q$  is the number of quantization levels. In this region, which is known as the overload region, the error increases monotonically with the magnitude of the input signal without any limits.

The second region in Figure 7.13 is when the input signal is small in magnitude ( $|v_{in}| \leq \frac{Q}{2}\Delta$ ). In this region, which is known as the granular noise region, the quantization error oscillates between  $\pm \frac{\Delta}{2}$ .

Assume that the input signal is such that the quantizer doesn't go to overload. Furthermore, assume that the quantization error is a random variable uniformly distributed between  $\Delta/2$  and  $-\Delta/2$ , and is independent of the analog input. Hence, the probability of the quantization error is as shown in Figure 7.14, and is given by the following equation:

$$P(q) = \begin{cases} \frac{1}{\Delta} & -\frac{\Delta}{2} < q < \frac{\Delta}{2} \\ 0 & \text{Otherwise} \end{cases} \quad (7.27)$$

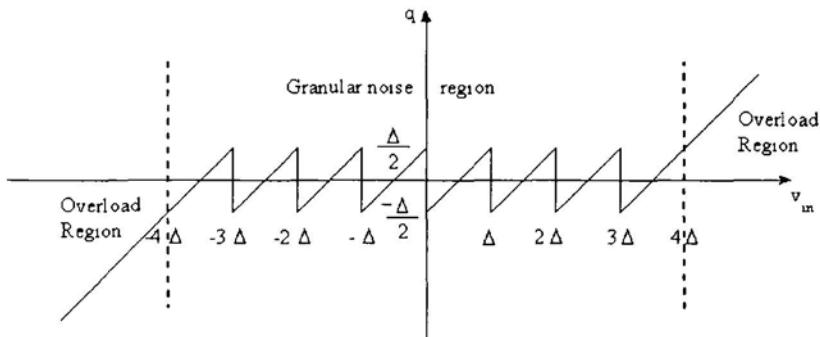


Figure 7.13. Quantization error versus input signal's magnitude, for an eight-quantization level mid-tread quantizer.

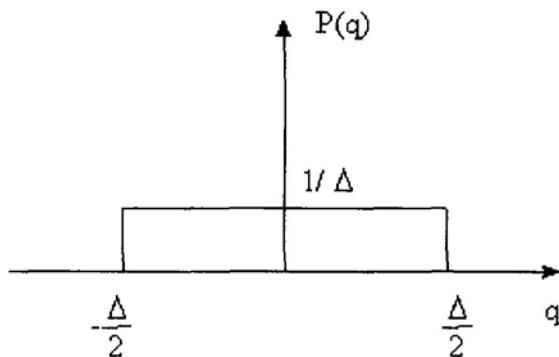


Figure 7.14. Probability density function of the quantization error.

Therefore, the quantization noise  $N_q$  is given by:

$$\begin{aligned}
 N_q &= \int_{-\frac{\Delta}{2}}^{\frac{\Delta}{2}} \frac{1}{\Delta} q^2 dq \\
 &= \frac{\Delta^2}{12}
 \end{aligned} \tag{7.28}$$

If,  $V_{PP}$  is the peak-to-peak value of the input signal. Then,

$$\Delta = \frac{V_{PP}}{Q} = \frac{V_{PP}}{2^n} \quad (7.29)$$

Where,  $n$  is the number of bits at the output of the analog-to-digital converter. Therefore,

$$N_q = \frac{1}{12} \frac{V_{PP}^2}{2^{2n}} \quad (7.30)$$

Assuming that the input signal is a sinusoidal wave, hence the input signal power is given by:

$$S = \frac{V_{PP}^2}{8} \quad (7.31)$$

Therefore, the input signal-to-quantization noise ratio is given by:

$$\begin{aligned} \text{SNR}_q &= \frac{3}{2} 2^{2n} \\ &= 1.77 + 6n \text{ (in dB)} \end{aligned} \quad (7.32)$$

A more detailed analysis of quantization is given in [44].

## 7.6 TYPES OF ANALOG-TO-DIGITAL CONVERTERS

Analog-to-digital converters are classified into one-step architectures, such as: flash, folding and interpolative topologies, and multi-step architectures, such as: Successive approximation and pipeline topologies.

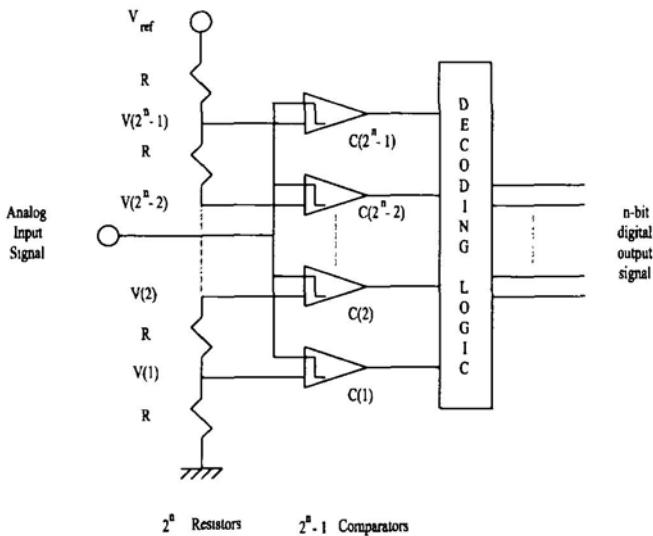
### 7.6.1 Flash Analog-to-Digital Converter

The flash architecture is potentially the fastest analog-to-digital converter, because of the parallelism it employs. However, this parallelism puts a practical limit on the resolution of the flash analog-to-digital converter. For an  $n$ -bit analog-to-digital converter, the flash architecture employs  $(2^n - 1)$  comparators and decoding logic. A resistive ladder consisting of  $2^n$  equal resistors divides the reference voltage into  $2^n$  equally spaced voltages as shown in Figure 7.15

The output of the comparator constitutes a thermometer code. Assume that the analog input signal  $v_{in}$  is such that:

$$V_i < v_{in} \leq V_{i+1} \quad (7.33)$$

Therefore, the output of comparators  $C_1, C_2, \dots, C_i$  is logic 1, while the output of comparator  $C_{i+1}, C_{i+2}, \dots, C_{2^n-2}, C_{2^n-1}$  is logic 0. The decoding logic transfers the thermometer code which has  $(2^n - 1)$  bits into a binary



*Figure 7.15.* Block diagram of an  $n$ -bit flash analog-to-digital converter.

signal which has  $n$  bits Table 7.1 shows the mapping between the thermometer code and the binary coded signal when  $n = 3$ .

*Table 7.1.* The thermometer code of a binary coded signal for  $n = 3$ .

Thermometer code	Binary code
0000000	000
0000001	001
0000011	010
0000111	011
0001111	100
0011111	101
0111111	110
1111111	111

Flash architectures have the advantage of being very fast due to parallelism. State-of-the-art flash analog-to-digital converters achieve sampling rates as high as 300 MSamples/sec. However, they suffer a few disadvantages. First, due to the parallelism inherent in this architecture, the number of comparators increases exponentially with the resolution of the analog-to-digital converter. This places a limit on the practical resolution that can be achieved using the flash architecture to about 8 to 10 bits of resolution.

Second, flash architectures dissipate large power and use large area, especially as the resolution gets larger. Third, the input capacitance grows exponentially with the resolution of the analog-to-digital converter. Finally, the large number of comparators loads the ladder network, which causes a deviation in the voltage levels it generates.

### 7.6.2 Interpolative Analog-to-Digital Converters

Interpolative architectures are single-step analog-to-digital converters that have a lower input capacitance, area and power dissipation than the corresponding flash architecture that has the same resolution.

To illustrate the basic principle of an interpolative analog-to-digital converter, let's consider an architecture with an interpolation factor of 2, that has four quantization intervals. An analog-to-digital converter with four quantization intervals has three reference voltages;  $V_{r1}$ ,  $V_{r2}$  and  $V_{r3}$  that are related by the following equation:

$$V_{r3} - V_{r2} = V_{r2} - V_{r1} = \Delta \quad (7.34)$$

Where,  $\Delta$  is the step size of the analog-to-digital converter. Therefore,

$$V_{r2} = \frac{V_{r1} + V_{r3}}{2} \quad (7.35)$$

i.e.,  $V_{r2}$  is the average (interpolation by a factor of 2) of its two adjacent reference voltages. This is the basic principle of the interpolative analog-to-digital converter. Consider the block diagram shown in Figure 7.17, it consists of two amplifiers having inverting and non-inverting outputs. The transfer characteristics of each amplifier is given by:

$$V_{out}^+ = \begin{cases} V_s^- & v_{in} - V_{ref} < \frac{V_s^- - V_{DC}}{A} \\ A(v_{in} - V_{ref}) + V_{DC} & \frac{V_s^- - V_{DC}}{A} < v_{in} - V_{ref} < \frac{V_s^+ - V_{DC}}{A} \\ V_s^+ & v_{in} - V_{ref} < \frac{V_s^+ - V_{DC}}{A} \end{cases} \quad (7.36)$$

$$V_{out}^+ = \begin{cases} V_s^+ & v_{in} - V_{ref} < -\frac{V_s^+ - V_{DC}}{A} \\ A(v_{in} - V_{ref}) + V_{DC} & -\frac{V_s^+ - V_{DC}}{A} < v_{in} - V_{ref} < -\frac{V_s^- - V_{DC}}{A} \\ V_s^- & v_{in} - V_{ref} < -\frac{V_s^- - V_{DC}}{A} \end{cases} \quad (7.37)$$

Where

$A$  is the gain of the amplifier in its linear region of operation.

$V_{DC}$  is the DC offset at the output of the amplifier.

$V_S^+$  is the upper saturation level at the output of the amplifier.

$V_S^-$  is the lower saturation level at the output of the amplifier.

$v_{in}$  is the input voltage on one input of the amplifier.

$V_{ref}$  is the reference input voltage on the other input of amplifier.

$V_{out}^+$  is the non-inverting amplifier output.

$V_{out}^-$  is the inverting amplifier output.

Figure 7.16 shows the transfer characteristics for inverting and non-inverting outputs of this amplifier.

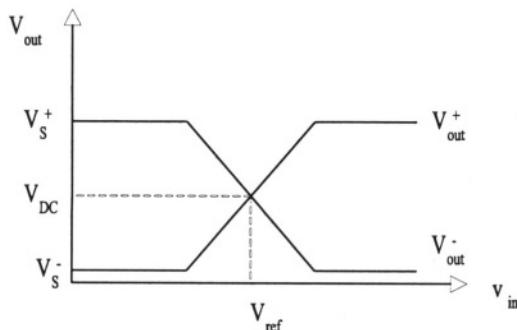


Figure 7.16. Transfer function of the amplifier used in the interpolative analog-to-digital converter.

In figure 7.17, the input reference voltage of the first amplifier is  $V_{r1}$ , and that of the second amplifier is  $V_{r3}$ . The inverting and non-inverting outputs of the first amplifier are  $V_1^-$  and  $V_1^+$  respectively. While those of the second amplifier are  $V_2^-$  and  $V_2^+$ . Therefore, assuming that both amplifiers are in the linear range of operation, we have:

$$V_1^+ = A(v_{in} - V_{r1}) + V_{DC} \quad (7.38)$$

$$V_1^- = -A(v_{in} - V_{r1}) + V_{DC} \quad (7.39)$$

$$V_2^+ = A(v_{in} - V_{r3}) + V_{DC} \quad (7.40)$$

$$V_2^- = -A(v_{in} - V_{r3}) + V_{DC} \quad (7.41)$$

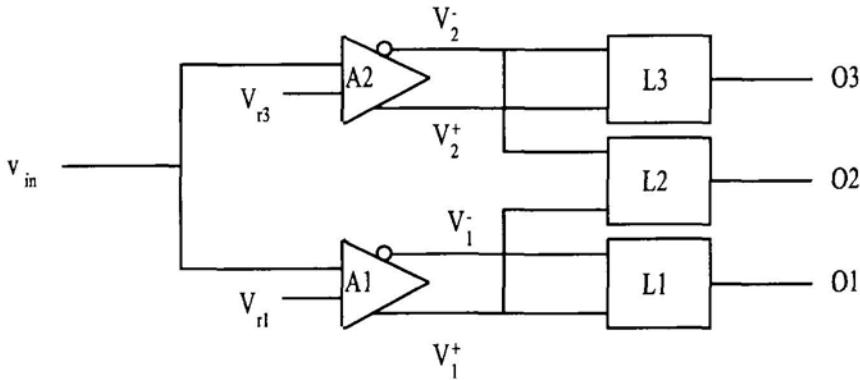


Figure 7.17. Interpolative analog-to-digital converter with an interpolation factor of 2.

Following the amplifiers in Figure 7.17 are three latches. Each latch has two inputs denoted by  $V_u$  and  $V_l$ , such that if  $V_u < V_l$ , the latch stores a one, and if  $V_l \leq V_u$ , the latch stores a zero.

For latch 1, if  $V_1^- \geq V_1^+$  ( $v_{in} \leq V_{r1}$ ), it stores a zero. But if  $V_1^+ > V_1^-$  ( $v_{in} > V_{r1}$ ), it stores a one. Similarly for latch 3, if  $V_2^- \geq V_2^+$  ( $v_{in} \leq V_{r3}$ ), it stores a zero. But if  $V_2^+ > V_2^-$  ( $v_{in} > V_{r3}$ ), it stores a one.

However, latch 3 compares the non-inverting output of amplifier 1 with the inverting output of amplifier 2. When  $V_2^- \geq V_1^+$ , it can be shown that  $v_{in} \leq \frac{V_{r1} + V_{r3}}{2}$ , or equivalently  $v_{in} \leq V_{r2}$ . In this case latch 2 stores a zero. On the other hand, if  $V_2^- < V_1^+$ , ( $v_{in} > V_{r2}$ ), latch 2 stores a one. Table 7.2 gives the output of each latch for each input voltage range.

Table 7.2. Latch outputs for an interpolative analog-to-digital converter with an interpolation factor of 2.

Input voltage range	$O_1$	$O_2$	$O_3$
$v_{in} \leq V_{r1}$	0	0	0
$V_{r1} < v_{in} \leq V_{r2}$	1	0	0
$V_{r2} < v_{in} \leq V_{r3}$	1	1	0
$v_{in} > V_{r3}$	1	1	1

Even though we only used two reference voltages in the block diagram of Figure 7.17, yet by interpolating the output of the first amplifier with that of

the second amplifier we were able to obtain a logic output ( $O_2$ ) that reflects the relative value of  $v_{in}$  with respect to the average reference voltage of the two amplifiers ( $V_{r2} = \frac{V_{r1} + V_{r3}}{2}$ ).

This interpolative analog-to-digital converter can be extended to a higher resolution (more quantization intervals) by using more amplifiers and latches. In general, an  $N$  quantization level analog-to-digital converter requires  $N/2$  amplifiers and  $N - 1$  latches. Notice that, the number of amplifiers is almost half that used in a flash analog-to-digital converter having the same resolution. This is how the reduction in the input capacitance and the saving in area and power dissipation is achieved.

The interpolative analog-to-digital converter can be extended to higher interpolation factors by using voltage dividers between the outputs of the amplifiers as shown in Figure 7.18. For an interpolative analog-to-digital converter having  $N$  quantization levels and interpolation factor of  $m$ , the number of amplifiers needed is:

$$\frac{N - 2}{m} + 1,$$

and the number of latches needed is  $N - 1$ .

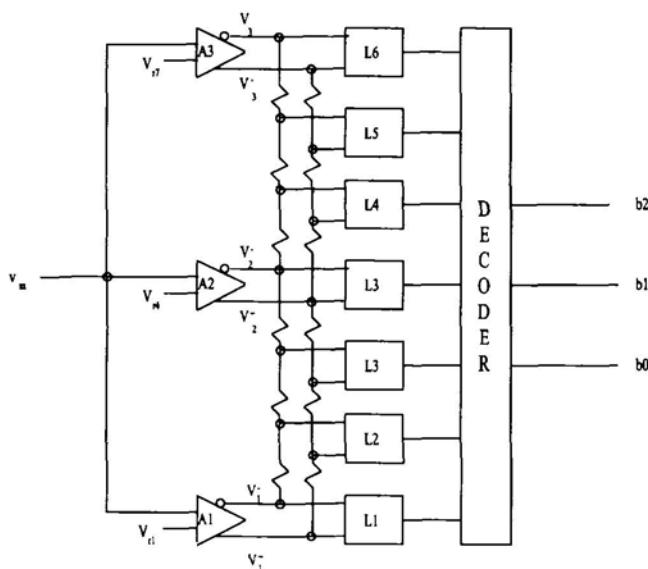


Figure 7.18. A three-bit interpolative analog-to-digital converter with an interpolation factor of three.

One potential problem for interpolative analog-to-digital converters is dead bands. Dead bands occur when the gain of the amplifiers used exceeds a certain value. Assume that  $V_{PP}$  is the peak-to-peak variation in the output voltage of the amplifier:

$$V_{PP} = V_S^+ - V_S^- \quad (7.42)$$

$\Delta$  is the input voltage range where the amplifier operates in the active region as shown in Figure 7.16. Therefore:

$$\Delta = \frac{V_{PP}}{A} \quad (7.43)$$

Consider now an interpolative analog-to-digital with an interpolation factor of 2, as shown in Figure 7.17. Figure 7.19 shows the relationship between  $(V_2^- - V_1^+)$  and  $v_{in}$  for three different values of  $A$ . When  $A \leq \frac{V_{PP}}{V_{r3}-V_{r1}}$ ,  $(V_2^- - V_1^+)$  is zero at a single point, (when  $v_{in} = V_{r2}$ ). On the other hand, if  $A > \frac{V_{PP}}{V_{r3}-V_{r1}}$ ,  $(V_2^- - V_1^+)$  is zero over a range of  $v_{in}$  given by:

$$V_{r1} \frac{V_{PP}}{2A} \leq v_{in} \leq V_{r3} - \frac{V_{PP}}{2A} \quad (7.44)$$

This region is known as the dead band. If the input signal values in this region, the polarity of  $(v_{in} - V_{r2})$  may not be properly detected by latch L2 of Figure 7.17. For this reason the gain of the amplifier needs to be limited to  $\frac{V_{PP}}{V_{r3}-V_{r1}}$ .

### 7.6.3 Two Step Analog-to-Digital Converters

Despite the high speed of flash architectures, their area, input capacitance, cost and power dissipation increases exponentially with resolution making them impractical to use for analog-to-digital converters with more than 8 to 10 bits of resolution. On the other hand, two step architectures trade speed for area and power dissipation.

Figure 7.20 shows a two step analog-to-digital converter. The first stage of the two step analog-to-digital converter finds a coarse estimate of the input analog signal. This generates the most significant bits of the digital output. The coarse estimate is converted back to the analog domain using a digital-to-analog converter and subtracted from the analog input. This difference, which is known as the residue is amplified and digitized by a second analog-to-digital converter, giving the least significant bits of the digital output.

A sample and hold circuit is essential at the input to the two-step analog-to-digital converter to keep the voltage constant during digitization. This is necessary because of the non-zero delay of the first stage (the coarse analog-to-digital converter) and the desire to subtract the estimated voltage from the analog input that generated this estimate.

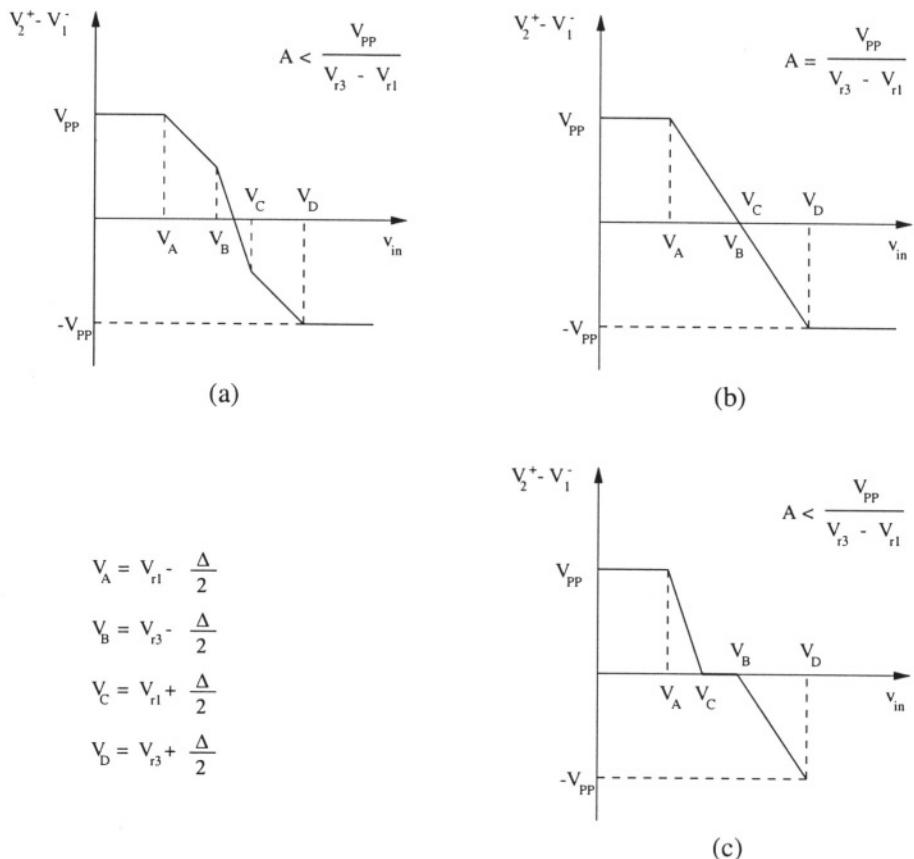


Figure 7.19. Relation between  $(V_2^+ - V_1^-)$  and  $v_{in}$  for different values of the amplifier gain ( $A$ ), for the interpolative analog-to-digital converter of Figure 7.17.

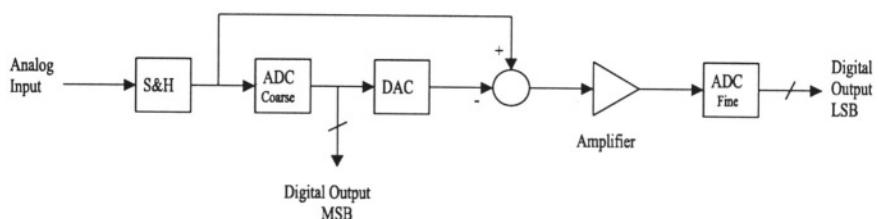


Figure 7.20. A two-step analog-to-digital converter.

The residue generated by the subtractor is a small signal smaller than the step size of the coarse analog-to-digital converter. If this signal were to be digitized directly by the second analog-to-digital converter, it would need a high precision analog-to-digital converter. Another option would be to use an amplifier, as shown in Figure 7.20. Now that the residue has been amplified, a less precise analog-to-digital converter is used in the second stage. However, adding an amplifier has the disadvantage of increasing the overall delay of the architecture.

To increase the throughput of the two-step architecture, it is possible to pipeline the architecture by inserting a second sample and hold circuit after the amplifier and before the fine analog-to-digital converter of Figure 7.20, as shown in Figure 7.21. This allows each stage to process a different sample concurrently. Thus making the overall throughput dependent on the latency of the slowest stage and not on the overall latency of the architecture.

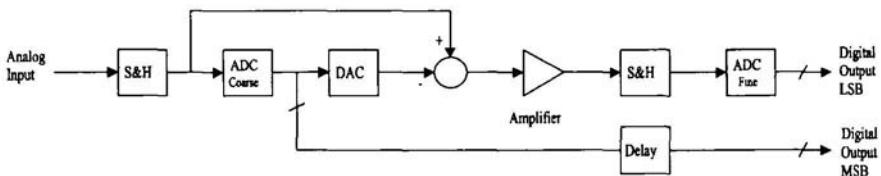


Figure 7.21. A two-stage pipelined analog-to-digital converter.

The operation of the two-stage pipelined analog-to-digital converter is quite simple. The first stage processors (coarse analog-to-digital converter, digital-to-analog converter, subtractor and amplifier) operates on the sample stored on the first sample and hold circuit. Upon completion, the result is passed to the second sample and hold circuit, while the first sample and hold circuit stores a new sample. The new sample is processed on by the first stage processors, while the old sample is being processed on by the fine analog-to-digital converter.

The above idea can be generalized to an  $n$ -stage pipelined analog-to-digital converter. Where, each of the first  $n - 1$  pipe-stages consist of a coarse analog-to-digital converter, a digital-to-analog converter, a subtractor and an amplifier, in addition to inter-stage sample and hold circuit. While, the last stage consists of a fine analog-to-digital converter. Each of the first  $n - 1$  pipe-stages generates a residue that is further digitized by the subsequent stages. The presence of the inter-stage sample and hold circuit allows each stage to process a different sample concurrently. It should be noted that the gain error and non-linearity of any stage needs to be less than the resolution

of the following stages. This places a high tolerance requirement on the gain linearity of the first few stages of the pipelined analog-to-digital converter.

The pipeline analog to digital converter is used for medium sampling frequencies in the range of 1 MSample/sec to 80 MSample/sec, and medium resolutions in the range of 8 to 12 bits.

A possible variant of the two-step architecture is the two step recycling architecture [45] shown in Figure 7.22. A single analog-to-digital conversion is used to perform both the coarse digitization of the input analog signal as well as the precise digitization of the residue.

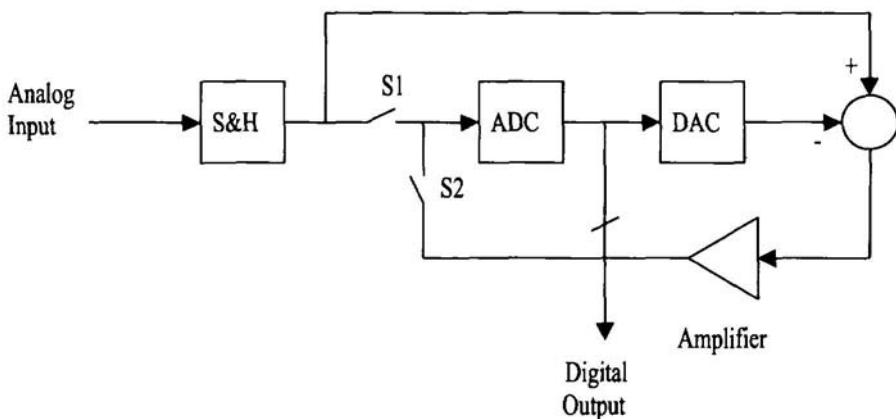


Figure 7.22. A two-step recycling analog-to-digital converter.

The theory of operation of the two step recycling analog-to-digital architecture is quite simple. During step-one, switch S1 is closed and switch S2 is open, the analog-to-digital converter digitizes the analog input signal. This digitization is a coarse digitization that generates the most significant bits of the digital output. The digital signal is converted back to the digital domain using a digital-to-analog converter, and is then subtracted from the analog signal, giving the residue which is then amplified.

During step-two, switch S1 is open and switch S2 is closed. Hence, the analog-to-digital converter digitizes the amplified residue generating the least significant bits of the digital output.

#### 7.6.4 Subranging Analog-to-Digital Converter

Another variant of the two step architecture is the subranging architecture [46]. The difference between this architecture and the conventional two-step analog-to-digital converter of Figure 7.20 is that this architecture doesn't use a subtractor, instead the output of the coarse analog-to-digital converter identifies

a narrow voltage range around the input signal. The second analog-to-digital converter, divides this narrow voltage into finer quantization levels that are used to determine the least significant bits of the digital output.

Figure 7.23 shows a block diagram of the subranging analog-to-digital converter. Assume that the coarse analog-to-digital converter is an  $m$ -bit analog-to-digital converter, and that the fine analog-to-digital converter is an  $n$ -bit analog-to-digital converter, then the total resolution of the subranging analog-to-digital converter is  $m + n$  bits.

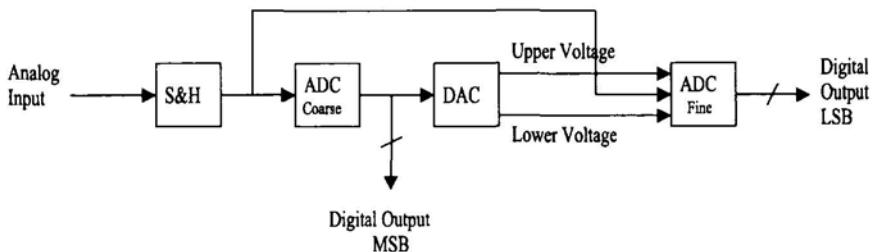


Figure 7.23. Block diagram of a subranging analog-to-digital converter.

Assume that the range of operation of the subranging analog-to-digital converter is  $[V_{max}, 0]$ . Hence, the lower and upper reference voltage level generated by the coarse analog-to-digital converter is:

$$R_L = 2^n \left\lfloor \frac{V_{max}}{2^n} \right\rfloor \quad (7.45)$$

$$\begin{aligned} R_U &= 2^n \left\lceil \frac{V_{max}}{2^n} \right\rceil \\ &= 2^n \left\lfloor \frac{V_{max}}{2^n} \right\rfloor + \frac{V_{max}}{2^n} \end{aligned} \quad (7.46)$$

Where,  $\lfloor \cdot \rfloor$  denotes the largest integer less than or equal to the argument, and  $\lceil \cdot \rceil$  denotes the smallest integer greater than or equal to the argument.

The coarse analog-to-digital converter subdivides the entire voltage range into  $2^m$  quantization intervals, as shown in Figure 7.24, such that if the input signal  $v_{in}$  is in the range:

$$\frac{i}{2^n} V_{max} \leq v_{in} < \frac{1+i}{2^n} V_{max}, \quad (7.47)$$

then the digital output of the coarse analog-to-digital converter is  $i$ , which represents the most significant  $m$  bits of the digital output of the subranging analog-to-digital converter.

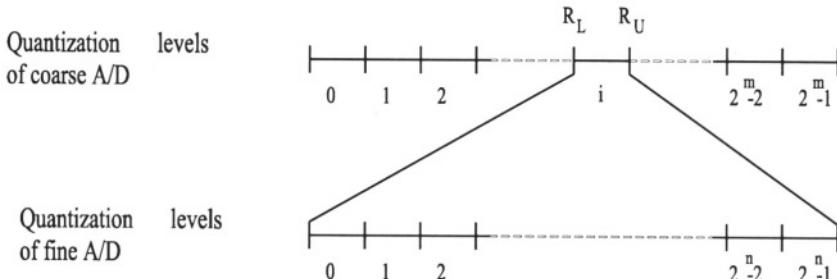


Figure 7.24. Quantization intervals in a subranging analog-to-digital converter.

The fine analog-to-digital converter operates over the narrower voltage range  $[\frac{i}{2^n}V_{max}, \frac{i+1}{2^n}V_{max}]$ . This voltage range is subdivided into  $2^n$  quantization intervals, as shown in Figure 7.24. This is used to determine the  $n$  least significant bits of the digital output.

### 7.6.5 Folding Analog-to-Digital Converters

Folding architectures are an evolution of two-step flash analog-to-digital converters. The basic principle of the folding architecture is to generate a residual voltage that is digitized with a fine analog-to-digital converter, to generate the least significant bits. A coarse analog-to-digital converter operates simultaneously on the analog input to generate the most significant bits.

Unlike the two-step architecture where a residual voltage is also generated, the generation of the residual voltage in the folding analog-to-digital converter doesn't depend on the coarse analog-to-digital converter nor does it require a digital-to-analog converter, rather it is generated by an analog folding circuit, whose theory of operation is explained in this section. This makes the folding analog-to-digital converter, shown in Figure 7.25, a one-step analog-to-digital converter.

To illustrate the principle of folding circuits consider the block diagram shown in Figure 7.26. Amplifier A1 has a gain of 1 in its active region, which extends between  $V_{r1}$  to  $V_{r2}$ . While amplifier A2 has a gain of -1 in its active region which extends between  $V_{r2}$  to  $V_{r3}$ . The output of the folding circuit is the sum of the outputs of the two amplifiers. The dependency of the output voltage of the folding circuit on the input voltage is as shown in Figure 7.26.

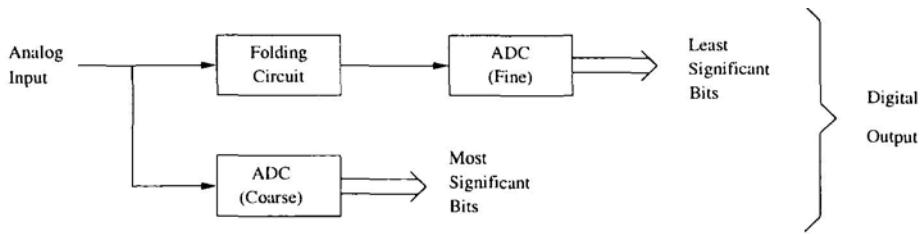


Figure 7.25. Block diagram of a folding analog-to-digital converter.

Below  $V_{r1}$  and above  $V_{r3}$ , the output voltage is zero. While between  $V_{r1}$  and  $V_{r3}$ , the output voltage increases linearly until it peaks at  $V_{r2}$ , where:

$$v_{out} = \Delta = V_{r2} - V_{r1} = V_{r3} - V_{r2}$$

The output voltage then starts decreasing linearly with the input voltage till it reaches zero when  $v_{in} = V_{r3}$ .

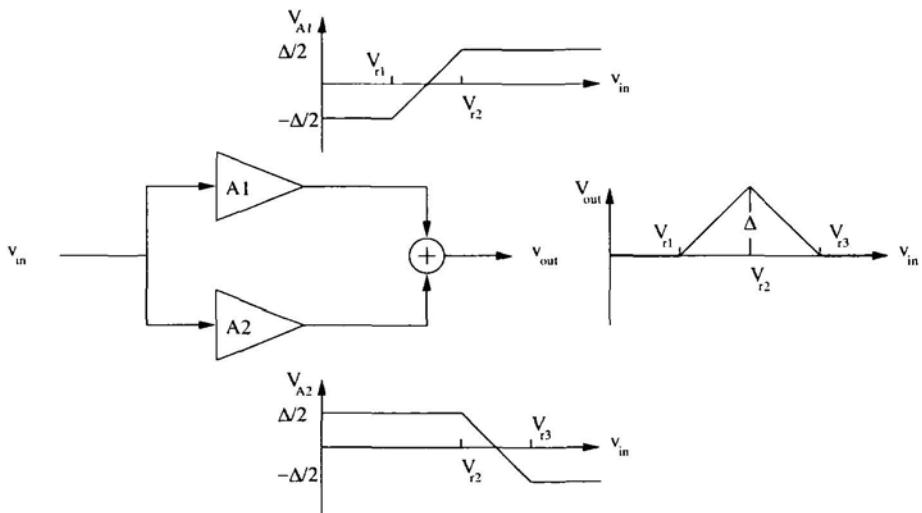


Figure 7.26. Block diagram of a folding circuit consisting of two amplifiers.

The reference voltages  $V_{r1}$ ,  $V_{r2}$  and  $V_{r3}$  determine the quantization interval of the coarse analog-to-digital converter. If the input voltage lies in the quantization interval  $[V_{r1}, V_{r2}]$ , the output voltage of the folding circuit is then given by:

$$v_{out} = v_{in} - V_{r1} \quad (7.48)$$

$v_{out}$  in this case represents the residual voltage. On the other hand, if the input voltage lies in the quantization interval  $[V_{r2}, V_{r3}]$ , the output voltage of the folding circuit is given by:

$$\begin{aligned} v_{out} &= V_{r3} - v_{in} \\ &= \Delta - (V_{in} - V_{r2}) \end{aligned} \quad (7.49)$$

In this case,  $v_{out}$  is related to the residual voltage by a sign inversion and a level shift. If we take these into account when we digitize the output voltage, it is possible to obtain the least significant bits of the digital output.

The folding circuit shown in Figure 7.26 can be used with a coarse analog-to-digital converter that has two quantization intervals. If the coarse analog-to-digital converter has more than two quantization levels, multiple folding circuits need to be used in parallel as shown in Figure 7.27, the outputs of these folding circuit is then added before being digitized by the fine analog-to-digital converter. Each folding circuit operates on a different range of the input voltage as shown in Figure 7.27.

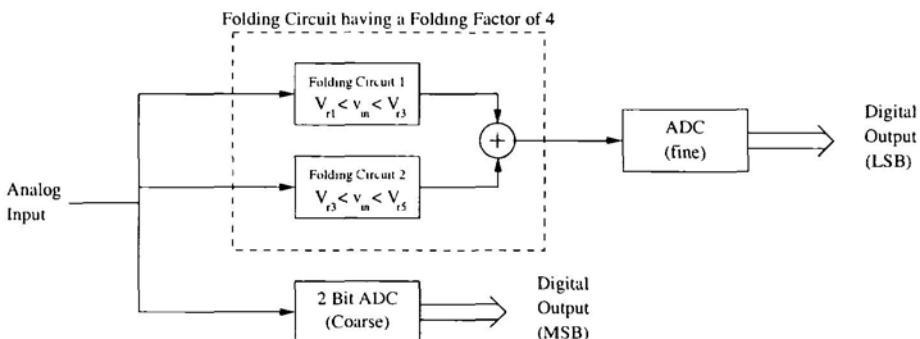


Figure 7.27. Block diagram of a folding analog-to-digital converter having a coarse analog-to-digital converter with 4 quantization intervals.

According to Figure 7.27 adding the outputs of two folding circuits with a folding factor of 2 and operating on adjacent voltage ranges gives a folding circuit with a folding factor of 4. Figure 7.28 shows the block diagram along with the transfer characteristics of such a folding circuit. In general, a folding circuit with a folding factor of  $N$  consists of  $N$  amplifiers having adjacent but non-overlapping active regions. The transfer characteristics of such a folding circuit has  $N/2$  peaks and  $N/2-1$  troughs.

Despite the simplicity of folding analog-to-digital converters relative to two-step analog-to-digital converters, folding analog-to-digital converters don't require subtractors, digital-to-analog converters nor sample and hold circuits, and despite the fact former are faster than the latter because of the parallel

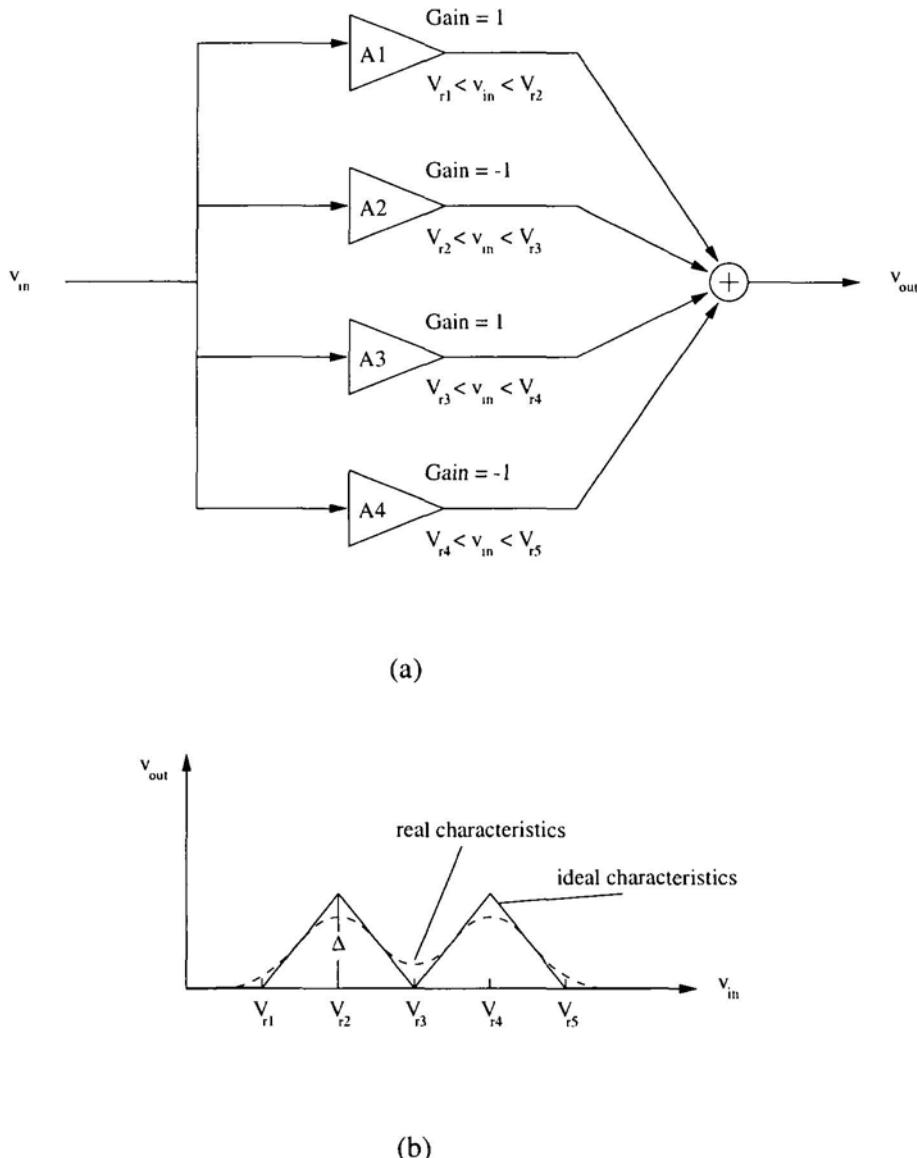


Figure 7.28. (a) Folding circuit with a folding factor of 4. (b) Transfer characteristics of such a folding circuit.

nature of its operation, yet, folding analog-to-digital converters suffer from a few drawbacks.

First, as the resolution of the coarse analog-to-digital converter increases, the complexity of the folding circuit grows exponentially. Second, the amplifiers

employed in the folding circuits were assumed to be ideal in the sense that, the gain is unity in the active region and zero in cutoff and the transition between the active region and cutoff is a sharp transition. In practice, this is not the case, the gain decreases gradually from the active region to cutoff, making the transfer characteristics of the folding circuit some what smooth as shown in Figure 7.28.b.

Finally, the fact that folding circuit has so many peaks and troughs necessitates that the folding circuit should have a high frequency of operation. For example, consider a folding circuit with a folding factor  $N$ , assume that the input voltage changes suddenly from maximum value to minimum value, the output of the folding circuits goes through  $N/2$  peaks, resulting in a higher frequency that the folding circuit needs to handle.

### **7.6.6 Successive Approximation Analog-to-Digital Converters**

Successive approximation is a multi-step analog-to-digital architecture, with the number of steps being equal to the resolution of the analog-to-digital converter. This converter uses a binary search algorithm, where the search interval is narrowed by half during each step by comparing the analog input with the mid-point of the search interval. If the analog input is larger, the corresponding bit of the digital output is set to one, and the upper half of the current search interval is selected as the search interval during the next step. If the analog input is smaller than the mid-point of the search interval, the corresponding bit of the digital output is set to zero, and the lower half of the current search interval is selected as the search interval during the next step.

Initially, the search interval is set to the entire voltage range of the analog-to-digital converter. As we proceed from one step to the next we select the most significant bits to the least significant bits. During each step the search interval is half the search interval of the previous step. Figure 7.29 shows the output waveform of a four step analog-to-digital converter generating a four digital signal.

Figure 7.30 shows the block diagram of a successive approximation analog-to-digital converters. The digital word stored in the register corresponds to the middle of the current search interval. Initially, a one stored is stored in the most significant bit and a zero in all other bits. The digital-to-analog converter (DAC) converts this digital signal into an analog signal that is compared with the input analog signal. The output of the comparator goes to the control unit. If the control unit determines that the analog input exceeds the output of the DAC, the “one” is retained in MSB, and a “one” is stored in the next bit, otherwise the MSB is reset to “zero” and a “one” is stored in the next bit. The output of the DAC corresponds to the mid-point of the new search

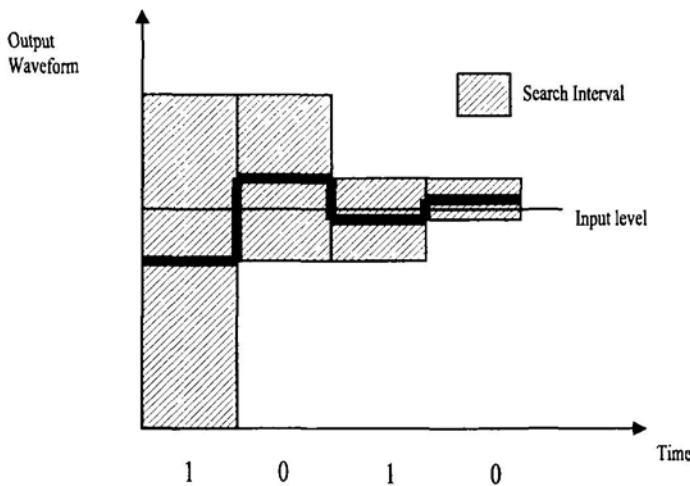


Figure 7.29. Output waveform of a 4-bit successive approximation analog-to-digital converter.

interval. This process continues for  $N$ -step until the  $N$ -bits of the digital output are generated.

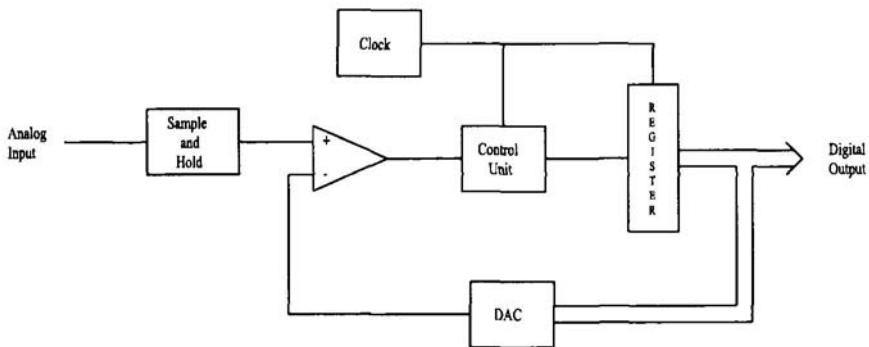


Figure 7.30. Block diagram of a successive approximation analog-to-digital converter.

For an  $N$ -bit digital output, the successive approximation analog-to-digital converter requires  $N$  clock cycles, making the successive approximation architecture about  $N$  times slower than the flash architecture. The maximum allowed clock frequency is limited by several factors, such as the setup and

hold times of the register, the settling time of the digital-to-analog converter and the delay of the comparator. Despite its slow speed, the successive approximation architecture has a few advantages such as its simplicity and its ability to produce a high resolution digital output.

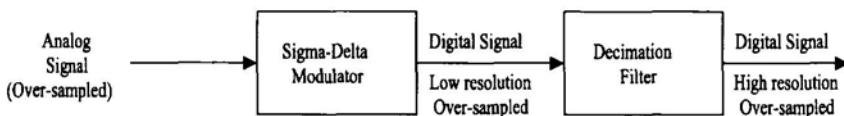
During analog-to-digital conversion, the input analog signal to the comparator needs to be held constant. This necessitates the use of a front end sample and hold circuit before the comparator, as shown in Figure 7.30.

Successive approximation analog-to-digital converters achieve resolutions of up to 14 bits. However, they suffer from their slow sampling speeds, which is limited to 1 MSample/sec. The successive approximation analog-to-digital converter is inherently programmable, allowing the designer to trade resolution for speed. Successive approximation analog-to-digital converters use a single comparator, this leads to very small die size and low power consumption.

## 7.7 SIGMA-DELTA ANALOG-TO-DIGITAL CONVERTERS

Sigma-Delta analog-to-digital converters are commonly used when high resolution is required, this is because of the sigma-delta modulator's ability to shape noise away from the desired band. Moreover, sigma-delta modulators require a two-level (one-bit) quantizer to achieve a high-resolution Nyquist rate sampled stream.

Figure 7.31 shows a block diagram of the Sigma-Delta analog-to-digital converter. It consists of two parts: A Sigma-Delta modulator followed by a decimation filter. The output of the sigma-delta modulator is a one-bit over-sampled stream. A decimation filter converts this over-sampled stream into a Nyquist rate sampled stream having high resolution.



*Figure 7.31. A Sigma-Delta Analog-to-Digital Converter.*

One distinct advantage of the Sigma-Delta technique is that analog signals are digitized using a one-bit quantizer which has a precision much less than the overall resolution of the analog-to-digital converter. The Sigma-Delta analog-to-digital converter is an over-sampling converter. The over-sampled

analog signal passes through a Sigma-Delta modulator that quantizes the signal to one bit, sometimes a two-bit or higher resolution quantizer is used.

The Sigma-Delta modulator shapes the noise away from the desired signal band, such that when the signal is decimated to the Nyquist rate, the in-band quantization noise is small, which in turn leads to a high resolution analog-to-digital converter. The effectiveness of the Sigma-Delta modulator in removing the quantization noise from the desired signal band depends on the order of the Sigma-Delta modulator as well as the over sample factor. The higher the order of the Sigma-Delta modulator and the higher the over-sampling factor the lesser the in-band quantization noise and the higher the resolution at the output of the Sigma-Delta analog-to-digital converter.

Figure 7.32.a shows a block diagram of a first-order Sigma-Delta modulator. The quantizer (comparator) can be modeled as a source of Additive White Gaussian Noise (AWGN). If we treat the quantization noise as being uniformly distributed over the range  $\pm\Delta/2$ . Where,  $\Delta$  is the step size of the quantizer. Then, the quantization noise has a mean square given by:

$$e_q^2 = \frac{\Delta^2}{12} \quad (7.50)$$

Hence, the Sigma-Delta modulator can be modeled as linear time invariant system with two inputs,  $e_s$  and  $e_q$  and one output  $e_o$ , as shown in Figure 7.32.b.

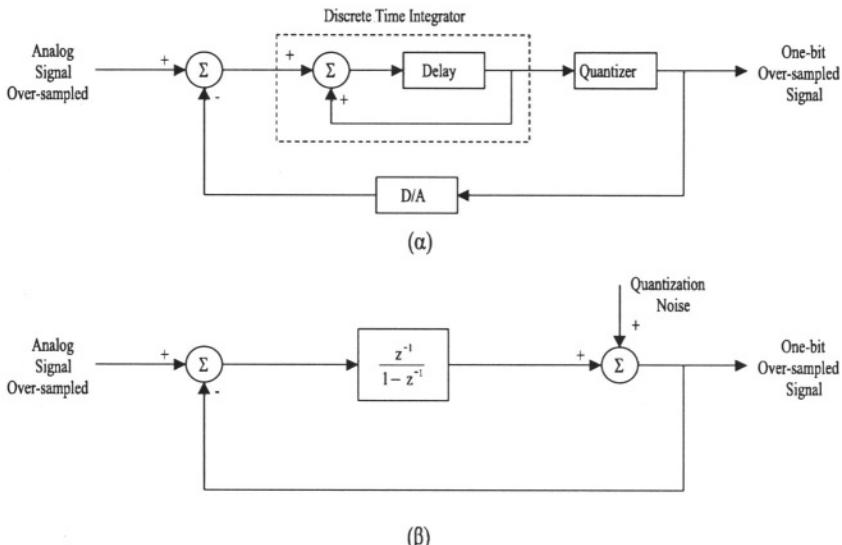


Figure 7.32. (a) Block diagram of a first-order Sigma-Delta modulator. (b) Linearized model.

The discrete time integrator of Figure 7.32.a has a transfer function given by:

$$\frac{z^{-1}}{1 - z^{-1}} \quad (7.51)$$

The output of the first-order sigma-delta modulator is given by:

$$e_o = z^{-1}e_s + (1 - z^{-1})e_q \quad (7.52)$$

Thus, the output signal of the sigma-delta modulator is the sum of two components, a delayed version of the input signal and a component due to the quantization noise. The quantization noise is shaped by the factor  $(1 - z^{-1})$  which a frequency domain response given by:

$$2 \sin\left(\frac{\omega T_s}{2}\right) \quad (7.53)$$

Where,  $1/T_s$  is the sampling frequency. Notice that, the Sigma-Delta modulator acts as a low-pass filter for the quantization noise as shown in Figure 7.33. To select the desired signal the output of the Sigma-Delta modulator is filtered through a decimation filter. The quantization noise component at the output of the decimation filter is given by:

$$e'_q = e_q \frac{\pi}{\sqrt{3K^{3/2}}} \quad (7.54)$$

Where,  $K$  is the over-sampling factor. This is the ratio between the sampling frequency and the Nyquist frequency which is given by:

$$K = \frac{1}{T_s f_n} \quad (7.55)$$

Notice that, for the a first-order Sigma-Delta modulator, every time the sampling frequency increases by one octave ( $K$  doubles), the quantization noise at the output of the decimation filter decreases by 9 dB which corresponds to an increase in resolution by 1.5 bits.

The block diagram of the Sigma-Delta modulator can be modified to further limit the amount of quantization noise inside the desired signal band. Figure 7.34 shows a block diagram of a second-order Sigma-Delta modulator. The output of the second-order Sigma-Delta modulator is given by:

$$e_o = z^{-1}e_s + e_q(1 - z^{-1})^2 \quad (7.56)$$

The quantization noise is filtered by  $(1 - z^{-1})^2$  which attenuates the in-band quantization noise more than a first-order Sigma-Delta modulator. The in-band quantization noise for a second-order Sigma-Delta modulator is related to the overall quantization noise by the following equation:

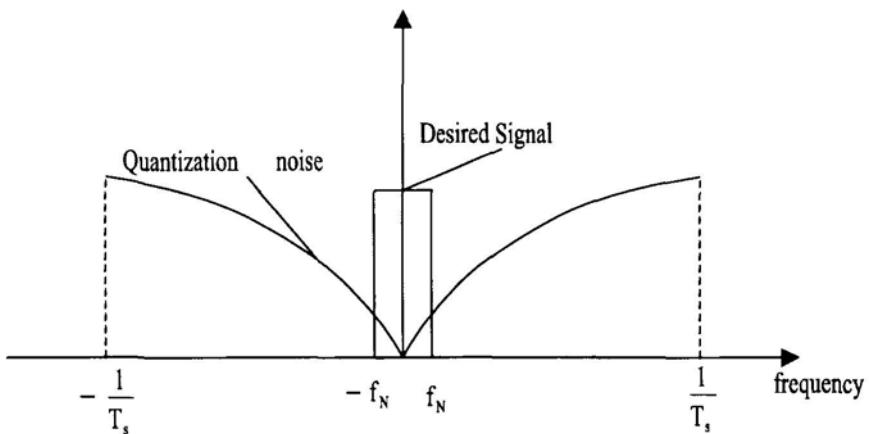


Figure 7.33. Frequency response of a first-order Sigma-Delta modulator.

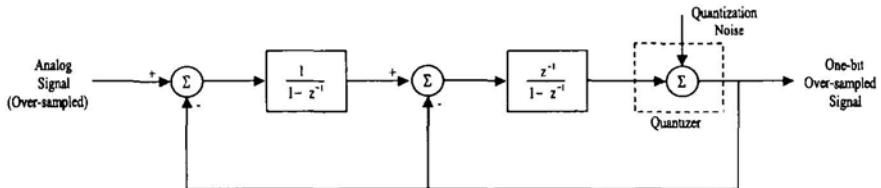


Figure 7.34. Block diagram of a second-order Sigma-Delta modulator.

$$e'_q = e_q \frac{\pi^2}{\sqrt{5} K^{5/2}} \quad (7.57)$$

In this case, every time the sampling frequency increases by one octave ( $K$  doubles), the in-band quantization noise decreases by 15 dB, which corresponds to an increase in the resolution of the analog-to-digital converter by 2.5 bits.

The order of the Sigma-Delta modulator can be extended to higher orders. In general, the output of the  $N$ th order Sigma-Delta modulator is given by:

$$e_o = z^{-1} e_s + e_q (1 - z^{-1})^N \quad (7.58)$$

The in-band quantization noise is given by:

$$e'_q = e_q \frac{\pi^n}{\sqrt{(2n+1)K^{(2n+1)/2}}} \quad (7.59)$$

In this case, every time the sampling frequency increases by one octave ( $K$  doubles), the in-band quantization noise decreases by  $(6n + 3)$  dB, which

*Table 7.3.* Dynamic range versus OSR for first, second and third order Sigma-Delta modulators.

OSR	1st Order	2nd Order	3rd Order
8	22 dB (3 bits)	32 dB (5 bits)	42 dB (7 bits)
16	31 dB (5 bits)	47 dB (7 bits)	63 dB (10 bits)
32	40 dB (6 bits)	62 dB (10 bits)	84 dB (14 bits)
64	49 dB (8 bits)	77 dB (12 bits)	105 dB (17 bits)
128	58 dB (9 bits)	92 dB (15 bits)	126 dB (21 bits)

corresponds to an increase in the resolution of the analog-to-digital converter by  $(n + 0.5)$  bits. Table 7.3 gives the output dynamic range of the Sigma-Delta modulator for different over-sampling ratios, for first-order, second-order and third-order Sigma-Delta modulators.

So far, we have assumed that the analog signal is a low-pass signal and that the sampling frequency is much larger than the Nyquist rate. Sigma-Delta modulators can also be used for the analog-to-digital conversion of band-pass signals if the signal bandwidth is much larger than the sampling frequency [47], [48], [49]. The sampling frequency of the band-pass signal is related to its carrier frequency and is given by:

$$f_s = \frac{4f_c}{2k - 1} \quad (7.60)$$

Where,  $k$  is a positive integer. Usually,  $k = 1$ . Figure 7.35 shows the block diagram of a second-order band-pass Sigma-Delta modulator. The dynamic range performance of a second-order Band-pass Sigma-Delta modulator is equivalent to that of a first-order Low-pass Sigma-Delta modulator. The output signal of the second-order Band-pass Sigma-Delta modulator is given by:

$$e_o = z^{-2}e_s + e_q(1 + z^{-2}) \quad (7.61)$$

The quantization noise transfer function  $1 + z^{-2}$  has a zero at quarter the sampling. If  $f_s = 4f_c$ , then the quantization noise is shaped away from the carrier frequency. This enables the band-pass Sigma-Delta modulator to be used in the analog-to-digital conversion of band-pass signals.

The output of Sigma-Delta modulator is an over-sampled low-resolution signal. A decimation filter is needed to down-sample the signal to the Nyquist rate. The decimation filter (Figure 7.36) consists of two parts; the Sinc decimator and the low-pass decimation filter (LPDF). The Sinc decimator is characterized by its simple structure, requiring only addition operations which makes it a power-efficient structure. The order of the Sinc decimator used depends on the order of the Sigma-Delta modulator, and is given by [50]:

$$\text{Order of Sinc} = \text{Order of LPSD} + 1 \quad (7.62)$$

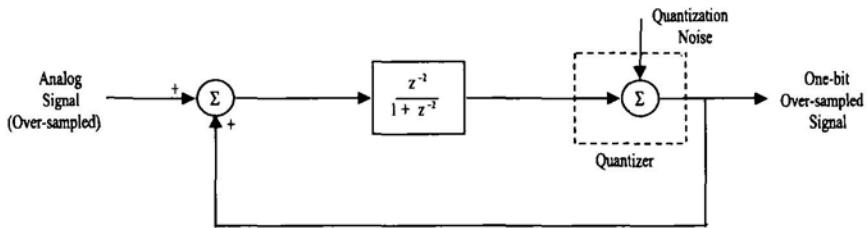


Figure 7.35. Block diagram of a second-order band-pass Sigma-Delta modulator.

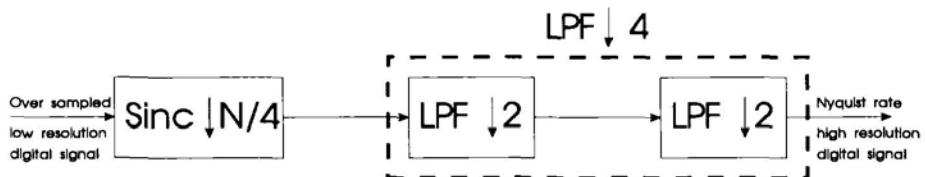


Figure 7.36. The decimation filter.

For a Sinc decimator with order  $N$  and a decimation factor of  $M$ , the transfer function (before down-sampling is given by):

$$H(e^{jw}) = (1 + e^{-jw} + e^{-2jw} + \dots + e^{-(M-1)jw})^N \quad (7.63)$$

Figure 7.37 shows the transfer function for a Sinc decimator having  $M = 8$ , and  $N = 3$ . Notice the zeros of the Sinc decimator at frequencies:  $\pm f_s/8$ ,  $\pm f_s/4$ ,  $\pm 3f_s/8$ , and  $\pm f_s/2$ . When the frequency spectrum is folded three times, around  $f_s/4$ ,  $f_s/8$ , and  $f_s/16$ , the zeros of the folded spectrum fall on the spectrum at  $f_s = 0$ . This minimizes the out-of-band noise added to the low frequency spectrum due to folding.

Due to its gradual transition from the pass-band to the stop-band, the Sinc decimator can't be used in the entire decimation process. The last stage of decimation is done using a low-pass decimation filter (LPDF), which does decimation by a factor of four [51]. The LPDF is built as a two stage LPDF each doing decimation by a factor of two.

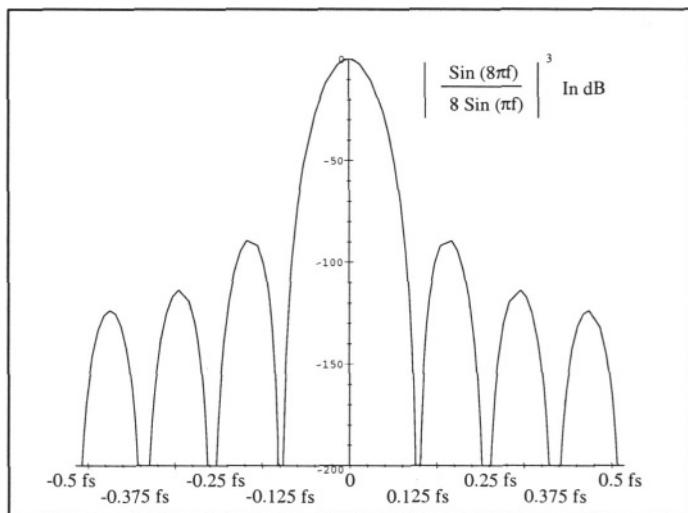


Figure 7.37. The transfer function of a Sinc decimator having,  $M = 8$ , and  $N = 3$ .

## Chapter 8

# VLSI DESIGN ISSUES IN WIRELESS TRANSCEIVER DESIGN

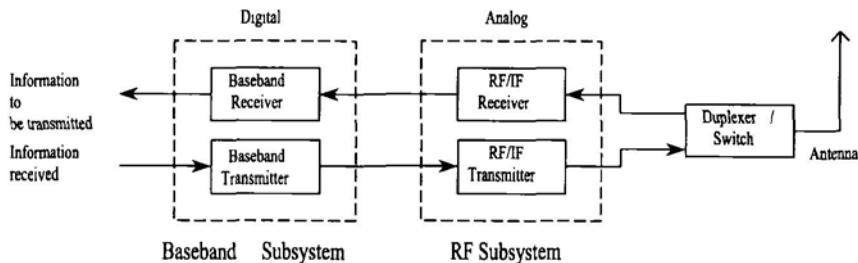
A typical portable mobile terminal uses several electronic components such as: analog-to-digital converters, digital-to-analog converters, ASICs, DSPs, as well as active and passive RF components. The choice of a certain architecture and accordingly, the type and number of components used depends on the design constraints imposed on the designer and on the acceptable system performance.

## 8.1 INTRODUCTION

The advancement in VLSI technology has dramatically impacted the development and advancement of wireless communication systems. On one hand, digital IC's are used for baseband and IF processing employing CMOS technology. While on the other hand, the front end RF portion of the transceiver is based on GaAs or Silicon Bipolar analog technologies. Silicon bipolar and/or CMOS technologies are employed in the intermediate amplifiers and up/down converters.

While IC technology will continue to play an important role in the affordability and versatility of wireless systems, it is important to focus on the current trends and endeavors in industry and academia that are going on to make this happen. The ultimate aim of these endeavors is to have a small, energy-efficient, low-cost and reliable mobile wireless system.

The signal processing functions performed by a wireless terminal can be classified into two subsystems as shown in Figure 8.1, the baseband subsystem and the radio frequency (RF) subsystem. Each subsystem has a different signal processing task which is reflected in the underlying architecture and the implementation technology.



*Figure 8.1.* Partitioning of a wireless mobile terminal.

The baseband subsystem operates on low frequency signals, with the signal processing predominately done in the digital domain. The purpose of this subsystem, on the transmitter side, is to convert the input signal, which is usually voice or low rate data, into a signal suitable for modulation, and transfer over a noisy wireless channel shared by other users. On the receiver side, the purpose of the baseband subsystem is to recover the information signal (voice or low rate data) from the noise and interference corrupted received data stream.

The RF subsystem operates on high frequency signals, in the IF and RF frequency bands, with the signal processing done in the analog domain. The purpose of this subsystem on the transmitter side is to up-convert the modulated baseband signal to the desired RF frequency as specified by the standard. This signal is also filtered and power amplified and then transmitted. On the receiver side, the RF subsystem amplifies the received signal, which could be as low as -120 dBm, down-converts it to baseband, and eliminates the adjacent channel interference through filtering.

Present day wireless mobile terminals are evolving from a three or four chip-set solution, with some external components, to a two chip-set solution, where one chip is for baseband digital signal processing and the other is for RF analog signal processing, with even fewer external components. A further step in the evolution of this design is the elimination of external components, which are required for filtering the RF analog signal, thus having a true integrated RF chip. This will in turn necessitate innovative design for the RF analog architecture.

Ultimately, the objective is to have a single-chip design for the wireless terminal, that integrates both the baseband digital circuits and the RF analog circuits on to a signal monolithic integrated circuit (system on chip) [52]. However, some issues still remain to be resolved for this to happen such as, the effect of the digital switching noise interference on the analog circuits. Furthermore, the technologies used for each of the baseband and RF subsystems

are incompatible, while high performance RF analog circuits are implemented in silicon bipolar or GaAs technologies, high density, power efficient baseband digital circuits are implemented in CMOS technology. This issue is being addressed in two different ways: First, using a BiCMOS technology where the bipolar part of this technology is predominately used for the RF analog circuits while the CMOS part is predominately used for the digital circuits. However, one drawback for using BiCMOS technology is that state-of-the-art BiCMOS processes have a larger feature size than state-of-the-art CMOS processes, which leads to chips having more area, higher cost, and higher power dissipation. Another approach is to use CMOS technology for the RF analog circuits, this is becoming possible as we go to deep sub-micron CMOS technology, which can operate at frequencies in the GHz range.

New techniques and algorithms in signal processing for wireless applications are emerging as the need to improve efficiency, lower power dissipation, and reduce cost continues to grow. In this chapter, we consider the implementation technology used for each of the baseband and RF subsystems to perform their tasks in a low-cost, power-efficient manner with acceptable performance.

In section 8.2, we discuss the design constraints imposed on wireless transceivers, we also consider some of the tradeoffs in achieving these constraints. In section 8.3, we present some of the devices used to implement the baseband subsystem and determine the advantages and disadvantages of each, we also consider the tradeoffs when selecting a device for the baseband subsystem. In section 8.4, we consider the RF subsystem, we present different architectures for the transmitter and receiver and consider the advantages and disadvantages of each. We also present the technologies competing for the implementation of the RF subsystem, and discuss issues related to some of the building blocks used in the RF subsystem.

## 8.2 TRANSCEIVER DESIGN CONSTRAINTS

The design of a portable mobile terminal involves many system, architectural, and technological tradeoffs to implement a cost viable, small size, light-weight and low-power portable transceiver with acceptable system performance.

Cost is one of the major design constraints that derives the design of a portable mobile terminal. About 50% of the terminal's cost goes into the baseband portion of the transceiver, 40% goes into the RF portion of the transceiver, while the remaining 10% are for the power supply. Cost is a driving factor in determining whether to implement certain functions, such as voice compression or channel coding, in hardwired ASIC hardware, or software programmable DSPs. Generally, a programmable architecture allows higher flexibility, but at greater cost (on the long run). As standards start to

evolve and become more mature, going to an ASIC might prove to be a more cost efficient solution.

Another factor influencing the over cost of the system is the level of integration. While baseband signal processing can be done using a few- or even one - ICs, RF signal processing requires a much higher number of components most of them being off-chip passive components (resistors, capacitors, inductors, etc.). Going to an integrated RF design at the present time might not be the most cost efficient design. However, as further progress goes on in the integrated RF circuit design arena this picture is expected to change.

Smaller size and lighter weight are two important goals for the design of a portable terminal. These can be greatly impacted by the level of integration. Going to higher component integration levels leads to smaller size and less weight. Furthermore, increasing the level of integration, reduces the off chip drive capacitance, which leads to lower power dissipation. Since the weight of the battery is a sizable portion of the mobile terminal's weight, reducing power dissipation leads to less terminal weight. Furthermore, reducing the power dissipation increases both the talk time and the stand-by time, which are two important consumer care-abouts.

The power amplifier is the dominant source of power dissipation in a portable mobile terminal, over 50% of the power dissipation can occur there. However, as the cell size continues to shrink to micro and pico cells, the transmit power will drop dramatically from a few watts to hundreds or even tens of milliwatts.

Power dissipation can be further reduced by voltage scaling, generally, power dissipation decreases with the square of the voltage, making voltage scaling an attractive method to reduce power dissipation. However, two problems start to appear as we reduce the voltage. First, the throughput of digital circuits drops in proportion with the decrease in voltage. Wireless systems are required to operate at a lower voltage while maintaining the overall system performance, this necessitates architectural modifications, such as pipelining and parallelism [53], to maintain the throughput as the voltage decreases.

Second, as the voltage decreases the dynamic range of the analog circuits decreases. It is conceivable that multiple supply voltages will be required for future portable terminals, with the digital subsystem using the lower voltage and the analog subsystem using the higher voltage. Powering down the unused portions of the transceiver is another effective way to further reduce the power dissipation. Because of its importance in the design of a wireless portable terminal, chapter 9, has been devoted to address the low power design issue.

After meeting all the design constraints the wireless system is required to achieve a certain level of system performance, such as the bit error rate

performance for a given signal-to-noise ratio (this is directly related to the voice quality), reliability of handoffs, talk time per unit battery weight, adjacent channel interference suppression, etc. The mobile terminal might also be required to provide more features and functionality such as data messaging and multi-standard, multi-band operation.

### 8.3 BASEBAND SUBSYSTEM DESIGN

The baseband subsystem performs most of its signal processing functions in the digital domain. In the receiver, the baseband analog signal coming out of the RF subsystem is digitized at the input to the baseband subsystem. It is also possible to digitize the signal at the IF stage [40]. The resolution of the analog-to-digital converter (ADC) depends on a variety of factors such as the number of channels digitized and the multiple access technique used. For example, digitizing a block of TDMA channels requires a high resolution ADC, with the resolution depending on the number of channels being digitized as well as adjacent channel interference suppression requirement. On the other hand, a CDMA receiver requires an ADC having a resolution as low as 4 bits.

In the transmitter, the digital signal is converted into an analog baseband (or IF signal) at the output of the baseband subsystem. The resolution of the digital-to-analog converter used depends on the dynamic range requirement of the transmitter. For example, to have a 100 dB dynamic range an 18 bit DAC is needed. However, this resolution may be reduced, if the gain of the analog transmitter amplifier is programable.

A combination of Application Specific Integrated Circuits (ASICs), Field Programmable Gate Arrays (FPGAs) and Digital Signal Processors (DSPs) are used in the baseband subsystem to perform the signal processing functions needed to produce the desired signal from the output of the analog-to-digital converter.

CMOS is the technology of choice in baseband to implement digital logic and memory VLSI circuits. Higher speed, lower power dissipation and higher circuit density (more integrability) are the critical considerations in choosing the baseband technology. To help achieve these objectives, the current trend is to scale the CMOS technology to sub  $0.1\mu m$ . The down-scaling of the CMOS technology will lead to further improvement in the integrability, size and power dissipation of the baseband digital subsystem.

A Digital Signal Processor (DSP) is essentially a microcomputer having a specialized hardware and instruction set that is tailored to do real time computational intensive digital signal processing functions. DSPs are highly flexible because they can be easily programmed to perform almost any digital signal processing algorithm using assembly language or high level languages such as C. DSPs allow fast design cycles thus reducing the time to market.

The most prevalent DSP architecture is the Harvard architecture [54]. This architecture is characterized by having separate data and instruction memories. DSPs differ from general purpose multiprocessors in several aspects. First, DSPs are customized to implement signal processing functions, such as finite impulse response (FIR) filters and fast Fourier transforms (FFTs). The core operation in these functions is the multiply-accumulate operation. Second, to be suitable for wireless applications, DSPs need to operate at lower power and be of lower cost than general purpose microprocessors.

The performance of a DSP is measured by how many million instructions it can execute per second, this is generally referred to as MIPS. Current CMOS technology supports DSPs operating at 50 to 100 MIPS. Lowering the power dissipation is essential in designing a mobile terminal. Hence, an important metric of the DSP is how much power it dissipates to execute one MIPS. This is generally expressed in mW/MIPS. As the CMOS technology is scaled down and the supply voltage is reduced, the power dissipation per MIPS is reduced.

In addition to the central processing unit, which is optimized to perform multiply-accumulate operations, the DSP has on chip memory both RAM and ROM, this is used to store both program instructions and data. Having on-chip memory avoids the delay penalty incurred when accessing off-chip memory. This is usually required to maintain the necessary throughput needed to implement signal processing algorithms in real time.

The current trend is to increase the on-chip memory, in order to increase speed (by reducing the memory access time), reduce power dissipation (by avoiding the power consuming off-chip memory access), enhance system performance and allow more functionality. DSPs can be used to implement controller operations, that don't take much of the DSP's horsepower, requiring only a few MIPS. However, controller operations take up substantial program memory. Another trend in DSP design is to integrate mixed-signal circuits, such as analog-to-digital and digital-to-analog converters, with the DSP core. This allows the integration of digital baseband and input/output functions on the same chip.

In a wireless systems, DSPs are used to implement real-time tasks. There are a few considerations that need to be taken into account when using a DSP to implement an algorithm:

1. The DSP selected should have an instruction set and architecture customized for wireless applications.
2. The speed of the clock should be selected as the minimum clock speed needed to implement the task in real time. This reduces the power dissipation by minimizing the time the DSP spends in the idle mode.
3. The DSP software needs to be properly architected to minimize the discrepancy between the peak MIPS load and the average MIPS load.

The DSP implements a signal processing function in a sequential manner. Consider, for example, an eight-tap FIR. To generate one output 8 multiply-accumulate operations are needed, these eight operations are executed sequentially. This reduces the data rate of the DSP. Furthermore, the datapath width of the DSP is fixed, making the DSP inefficient when it executes signal processing algorithms requiring low resolution. These problems can be avoided when using FPGAs or ASICs.

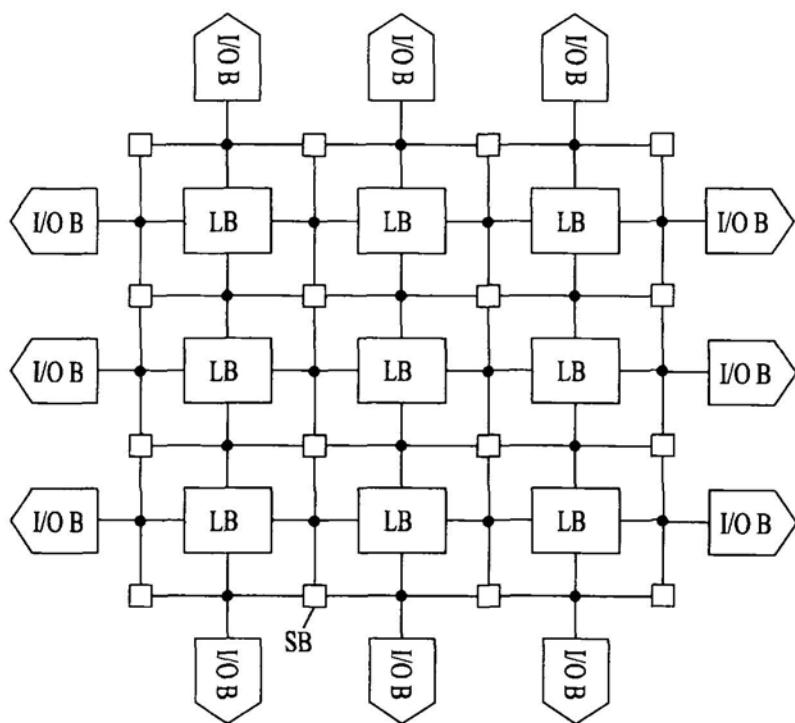
A Field Programmable Gate Array (FPGA) is an off-the-shelf VLSI device that allows the implementation of user-programmable designs. An FPGA consists of an array of logic blocks with an interconnection matrix in between. I/O blocks are used to connect the FPGA to the outside world. The interconnection matrix consists of horizontal and vertical wire segments and switching blocks. The switching blocks provide the connectivity between the vertical and horizontal wire segments. Figure 8.2 shows a block diagram of an FPGA.

The logic block of Figure 8.2 can vary in granularity anywhere from a single gate (such as a NAND gate) to more complex logic circuits having look up tables, multiplexers and flip flops.

A designer can configure the FPGA to perform almost any function by programming the logic blocks and switching matrix using the appropriate tools. The designer starts by describing his design using a schematic or in a hardware design language (HDL), such as Verilog or VHDL, or in a combination of both. HDLs are high level descriptions of the circuit behavior. This design is entered into a synthesis tool that maps the schematic or HDL description into the logic blocks and input/output blocks. The synthesis tool handles the low level details of connecting the logic blocks and input/output blocks according to the schematic or HDL description. Functionality and timing simulation are performed to verify functionality and compliance of the design with timing constraints. Finally, the file generated by the synthesis tool is downloaded to the FPGA. Figure 8.3 shows a diagram of the design flow.

The file generated by the synthesis tool is used to configure the FPGA. There are electric switches distributed throughout the FPGA, programming the FPGA involves programming these switches of the logic blocks to realize valid logic functions, and programming the switches of the interconnect matrix to realize valid wiring connections.

There are several types of programmable switches that can be used in the FPGA. The simplest switch is an RAM that controls a pass gate transistor, where according to the bit stored in the RAM cell the pass gate transistor can be conducting or non-conducting. The disadvantages of using this approach is that the design file needs to be downloaded to the FPGA every time it is powered up, because the contents of the RAM are volatile.



LB: Logic Block

SB: Switching Block

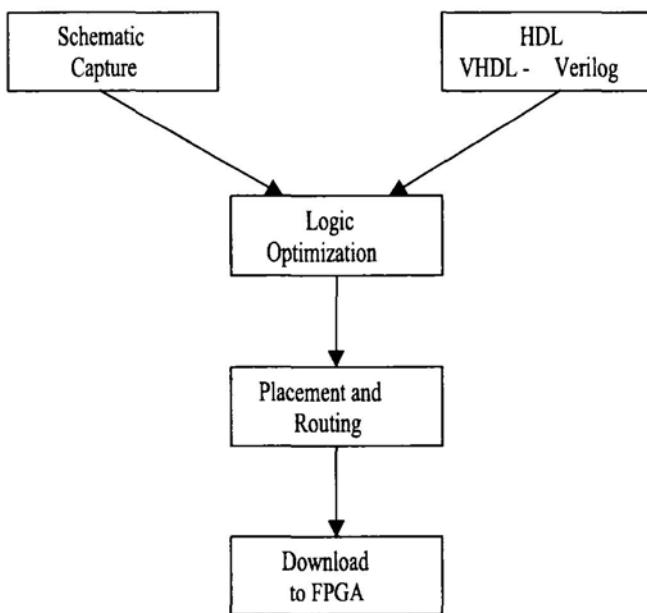
I/O B: Input/Output Block

*Figure 8.2. Block diagram of a Field Programmable Gate Array (FPGA).*

An alternative programming technology is to use Erasable Programmable Read Only Memory (EPROM) or Electrical Erasable Programmable Read Only Memory (EEPROM) transistors. The transistor is programmed by injecting a charge onto its floating gate to disable it. The state of the transistor is preserved until it is erased either by exposing it to ultra violet light, or electrically.

A third programming technology is to use antifuse devices. An antifuse is a high resistance device in its unprogrammed state. To program it, a high voltage is applied across its terminals, to get it burned and make its resistance low. Once the antifuse is burned, its resistance is permanently low and can't be made high again. The disadvantage of using antifuse programming technology is that the FPGA is programmed only once. Furthermore, to program the FPGA a special circuit is needed to generate the high voltage.

Using an FPGA has several advantages:



*Figure 8.3.* The FPGA design flow.

1. The design is reprogrammable except when using the antifuse programming technology. This flexibility allows the designer to respond quickly to an evolving standard. This is not possible with ASICs.
2. Higher degree of parallelism. Multiple execution units can be built to execute in parallel. On the other hand, a DSP has only one execution unit that executes operations sequentially.
3. It is possible to optimize the resolution of the datapath. This is not possible with DSPs.
4. Faster time to market when compared to ASIC designs.
5. A low risk design approach. The designer can quickly correct any design errors.
6. It is possible to implement reconfigurable logic.

However using FPGAs has its disadvantages:

1. Lower cost performance ratio than ASICs.

2. Lower circuit density and higher power dissipation than ASICs.
3. Longer design time than DSPs.
4. Special synthesis tools are needed to map the schematic or HDL description into the FPGA.

Application Specific Integrated Circuits (ASICs) are integrated circuits designed to implement a particular algorithm. When an ASIC is being designed, it is optimized for the algorithm it is implementing, this can lead to maximum resource saving, by avoiding resources that are required to support programmability. ASICs have the following advantages when compared to general purpose ICs (such as FPGAs and DSPs):

1. More power efficient.
2. Smaller size (less area).
3. Higher speed.

On the other hand ASICs require a longer design cycle and are more costly, for prototype products and products with small volume of production.

The choice of a particular type of integrated circuit (ASIC, FPGA, or DSP) depends on the algorithm being implemented. If the algorithm requires customized operations that are not suitably implemented on a DSP, it is implemented on an ASIC or an FPGA. On the other hand, if the algorithm being implemented is part of an evolving standard, then it is more suitable to implement the algorithm on a DSP or an FPGA. The choice of a particular type of integrated circuit also depends on the design-cycle time, with DSPs having the fastest design cycle time, while ASICs have the slowest design-cycle time. Power dissipation can also be another concern, with ASICs having the least power dissipation because of their custom design nature, while general purpose DSPs have the highest power consumption.

## **8.4 RF SUBSYSTEM DESIGN**

In a wireless system, the RF subsystem is located between the baseband subsystem and the antenna. The RF subsystem is divided into a transmit chain and a receive chain as shown in Figure 8.4. The transmit chain translates the frequency of the transmitted signal from baseband to the carrier frequency, which is in the range of 800 MHz to 2.5 GHz. The transmit chain is required to generate a signal with good linearity and high output power.

The receive chain of the RF subsystem is required to extract the desired signal from the background noise and interference. The receiver also translates this signal from the carrier frequency to baseband or IF (intermediate frequency),

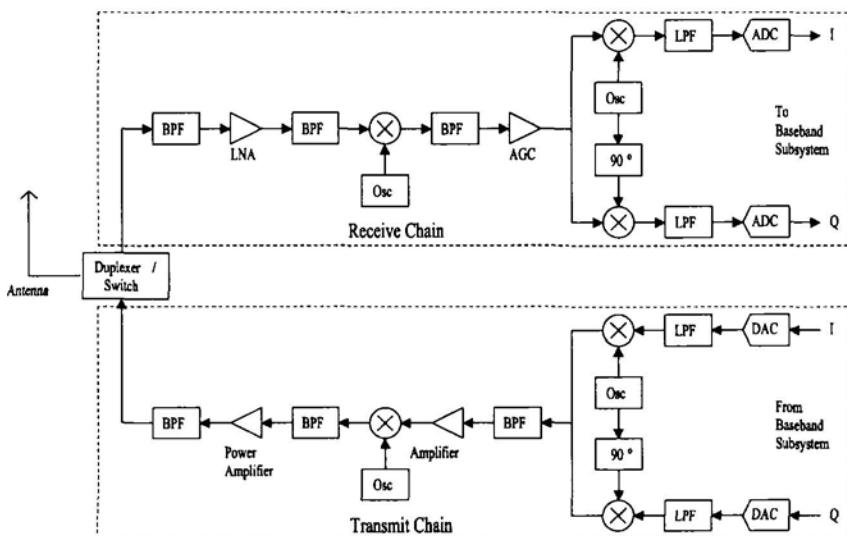


Figure 8.4. The RF subsystem of a wireless communications system.

where the signal is digitized and sent to the baseband subsystem for further processing.

The RF transceiver can be classified, according to the timing of the transmit and receive operations, into full duplex transceivers and half duplex transceivers. In a full duplex transceiver, the transmitter and receiver operate at the same time. On the other hand, in a half duplex transceiver, the transmitter and receiver operate at different times (the transmit and receive operations are multiplexed in the time domain). An example of a full duplex receiver is that used in IS-95, while the transceiver of GSM is a half-duplex transceiver. The isolation requirement, between the transmit and receive chains, is not as stringent in the half duplex transceivers as it is in the full duplex transceivers. This allows more integrability for half duplex transceivers.

The RF subsystem can also be classified according to the manner in which frequency translation between baseband and the carrier frequency occurs. A heterodyne receiver is one that converts the received signal to one or more intermediate (IF) frequencies, before eventually converting it to baseband. Similarly a heterodyne transmitter is one that up converts the transmitted signal to one or more IF frequencies, before eventually up converting it to the carrier frequency. The receiver and transmitter shown in Figure 8.4 are a heterodyne receiver and transmitter. On the other hand, a direct conversion

receiver (transmitter) is one that down (up) converts the signal from the carrier frequency (baseband) to baseband (the carrier frequency) directly.

Figure 8.5 shows a heterodyne receiver having one IF stage. The heterodyne receiver, when compared to a direct conversion receiver, has good dynamic range and sensitivity. However, its disadvantage is that it needs passive filters with high quality factors. The mixer of Figure 8.5 down converts both the desired signal (having a frequency  $f_{RF}$ ) and its image (having a frequency  $f_{RF} + 2f_{IF}$ ) to the IF frequency. To avoid mixing the image frequency to IF, the received RF signal is filtered before mixing as shown in Figure 8.5. Typically the IF frequency ( $f_{IF}$ ) is chosen to be one order of magnitude less than the RF frequency ( $f_{RF}$ ), this necessitates the use of a band-pass filter with a quality factor in the range of 50 to 100. Such a filter can't be integrated, instead discrete passive devices are used. High performance image rejection filters use ceramic resonators, while low-cost implementations use discrete inductor/capacitor filters. A band-pass filter is also needed at the IF stage of the receiver for selecting the desired channel. This filter is usually a surface acoustic wave (SAW) filter.

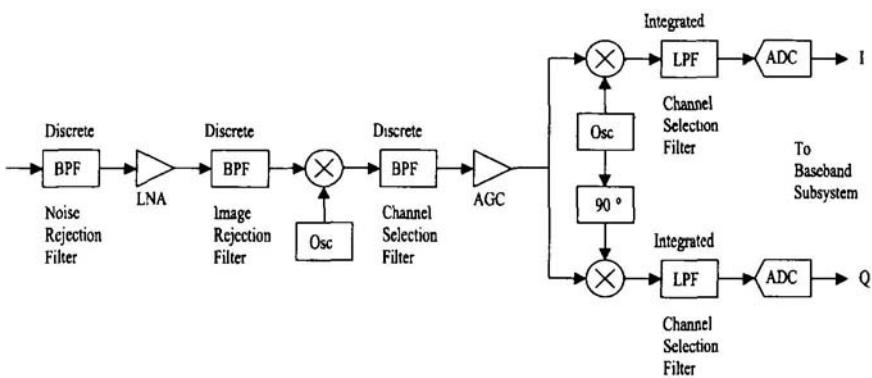


Figure 8.5. Single-stage heterodyne receiver.

Direct conversion receivers eliminate the need for the discrete band-pass filters used in heterodyne receivers, by converting the signal directly from the carrier frequency to baseband. The RF signal is down converted twice, once for the in-phase component and once for the quadrature-phase component. This type of down conversion is known as quadrature down conversion, and is shown in Figure 8.6.

Direct conversion receivers are easier to integrate than heterodyne receivers, this makes them more power efficient. The low-pass filters required after the

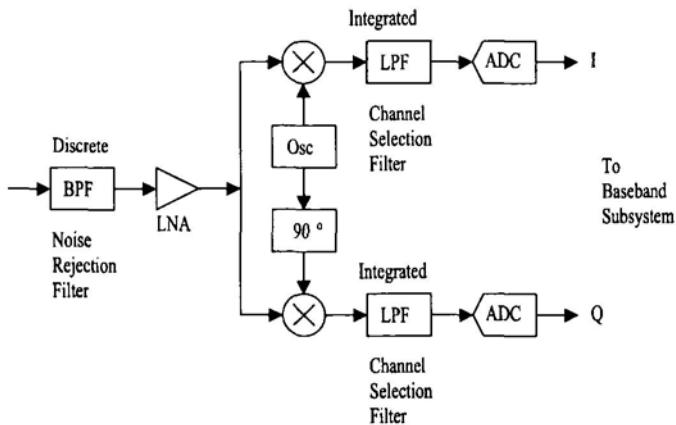


Figure 8.6. Direct conversion receiver.

mixers of the direct conversion receiver, can be easily implemented using integrated analog filters. However, a band-pass filter is still needed in the direct conversion receiver to reduce the dynamic range requirement of the mixer by filtering out the out-of-band noise. Band-pass filtering the RF signal also prevents mixing the RF signal with the harmonics of the oscillator. One challenge facing direct conversion receivers is the need for good matching between the in-phase and quadrature-phase branches. To get a 40 dB image rejection, the gain tolerance between the I and Q branches needs to be less than 1%.

A heterodyne transmitter converts the baseband signal to the carrier frequency through one or more IF stages. Figure 8.7 shows a heterodyne transmitter having a single IF stage. A band-pass IF filter is required to suppress the mirror signal during the up conversion process, this filter is also need to suppress any out-of-band spurious signals. Band-pass IF filters are implemented efficiently using off-the-chip discrete passive components, this prevents the integrability of the heterodyne transmitter.

A direct conversion transmitter, such as that shown in Figure 8.8, converts the transmitted signal directly from baseband to the carrier frequency, thus avoiding the need for off-chip filters. However, a filter is still needed at the output of the amplifier, to suppress any out-of-band spurious signals, to comply with the standard. The disadvantage of the direct conversion transmitter is the coupling that occurs between the output frequency and the local oscillator. Both these signals have the same frequency. Furthermore, the output of the power amplifier is a high-power signal, having a power as high as 1 W (30 dBm). This

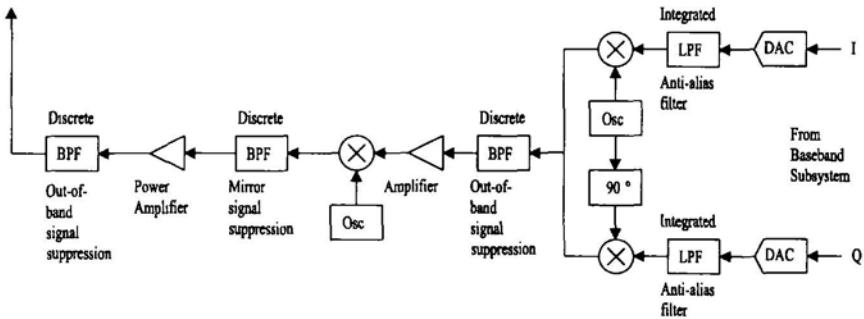


Figure 8.7. Heterodyne transmitter having a single IF stage.

signal can be easily coupled to the local oscillator resulting in undesirable spectrum broadening. One way to get around this problem is to use two local oscillators, multiply their signals together to get a signal having a frequency equal to the sum of the frequencies of the two local oscillators, this signal is used in the direct up conversion of the baseband signal to the carrier frequency. Figure 8.9 shows a block diagram of such a system, the drawback here being the need for an off-chip filter to suppress any spurious signals resulting from the multiplication of the signals of two oscillators together.

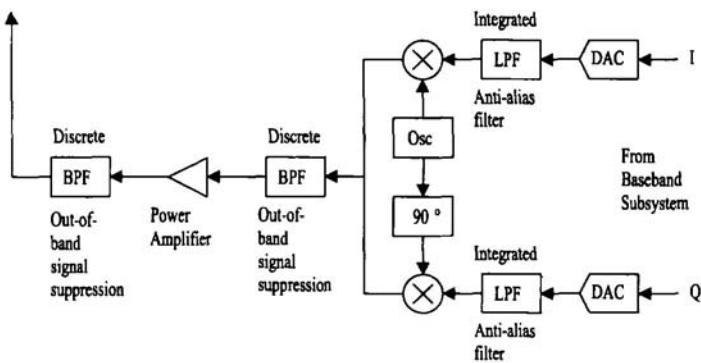


Figure 8.8. Direct conversion transmitter.

Several technologies compete for the implementation of the RF subsystem. These include GaAs, silicon bipolar, and bipolar CMOS (BiCMOS). Each of these technologies has its advantages and disadvantages, which influence the technology selection decision.

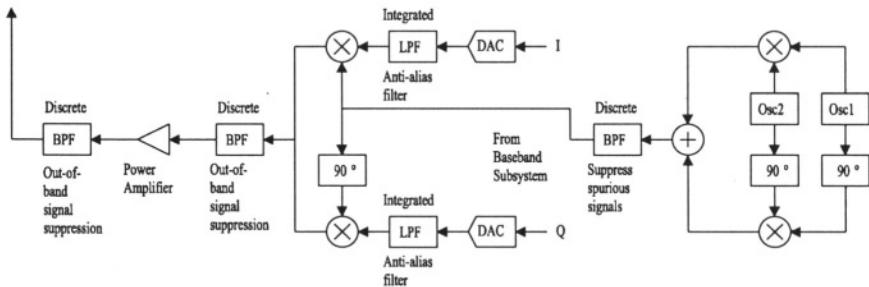


Figure 8.9. Direct conversion transmitter using two local oscillators.

The GaAs technology has the following advantages: High cutoff frequency - break down voltage product. High quality capacitors and inductors. Semi-insulating substrate. On the other hand, it also suffers some disadvantages: High power dissipation, a somewhat low yield, lower level of integration, and higher cost. The GaAs technology is used in the implementation of power amplifiers, front end switches and low noise amplifiers.

The silicon bipolar and BiCMOS technologies are typically used for the implementation of frequency synthesizers.

In addition to the three traditional technologies competing for the implementation of the RF subsection, a fourth competitor is the CMOS technology, which is the predominate technology used in the implementation of the baseband subsection. With the feature size continuing to shrink, as we go to deep sub-micron CMOS technology, the transit frequency is going up to tens of Giga Hertz.

However, to be a viable technology for the implementation of the RF subsystem, the CMOS technology needs to overcome the following obstacles [55]:

1. Substrate coupling of signals that differ in amplitude by up to 100 dB.
2. Parameter variation with process and temperature.
3. The need for accurate RF device models.

The RF subsystem consists of several electronic devices such as band-pass filters, low noise amplifiers (LNAs), automatic gain control (AGC) amplifiers, oscillators and mixers used for frequency up-conversion and frequency down-conversion, and power amplifiers.

The low noise amplifier (LNA) is the first stage of the receiver after the duplexer and the front-end filter. LNAs are designed to have a low noise figure in the range of 2 to 4 dB. The noise figure of the LNA is the dominant factor in determining the overall sensitivity (minimum detectable signal) of the receiver.

The LNA is also required to have high third-order intercept points, to minimize the in-band spurious components, when the interfering signals are strong. The LNA is also required to have high gain in the range of 10 to 16 dB, to minimize the noise contributions of the blocks following the LNA. The LNA is required to have low power consumption, because the receiver circuitry operates all the time. Typically, LNAs are designed to draw a current in the range of 2 to 6 mA. LNAs are usually designed to have a  $50\Omega$  input impedance and drive a  $50\Omega$  impedance.

To meet all the previous requirements and operate in the frequency range of 800 MHz to 2.5 GHz, several technologies can be used to implement LNAs. These technologies range from low-noise GaAs FETs, which are used to improve the front-end performance, to SiGe technology, Silicon Bipolar technology, BiCMOS technology, and even CMOS technology. State-of-the-art SiGe LNAs have noise figures as low as 1.3 dB in the frequency range 1.4 to 2.5 GHz, and dissipate as little as 10 mW and have a gain of 14 dB.

The mixer is a component used in both the transmit and receive chains of a transceiver for frequency up conversion and frequency down conversion respectively. Mixers can be implemented using diodes, Field Effect Transistors (FETs), or Bipolar Junction Transistors (BJTs). Diode mixers are passive mixers that introduce conversion loss. However, they have superior third order intercept performance than transistor mixers. Schottky diodes are frequently used in diode mixers. These diodes are characterized by their fast switching and small reactive parasitics making them suitable for high frequency applications.

Mixers using dual-gate MOSFETs or MESFETs<sup>1</sup>, have the RF signal and local oscillator signal applied to separate gates. This allows good isolation between the RF and local oscillator signals. Mixers using GaAs MESFETs can operate at high frequency and have low noise figures, because of the higher saturation velocity and higher mobility of electrons in GaAs. State-of-the-art GaAs MESFET mixers have gate length in the range of  $0.1\mu m$  to  $0.5\mu m$ . Silicon BJT mixers are usually implemented as Gilbert multipliers.

The Power amplifier is the last stage of the transmitter directly before the duplexer and the transmit filter. Power amplifiers are designed to have high linearity, this reduces spectrum re-growth. Power amplifiers are the most power consuming part of the system, hence they are required to have high efficiency. Furthermore, because of their high power consumption, power amplifiers are usually required to have a heat sink to prevent their overheating. Power amplifiers are commonly implemented in GaAs Microwave Monolithic Integrated Circuits (MMIC). State-of-the-art power amplifiers have a gain in the range of 26 dB and generate output power in the range of 30 dBm.

---

<sup>1</sup>A MESFET is a FET having a Schottky barrier diode across its gate

## Chapter 9

# LOW-POWER DESIGN TECHNIQUES

### 9.1 INTRODUCTION

In this chapter, the techniques used to lower the power dissipation for portable terminals in general and wireless portable terminals in particular are investigated. The last few years witnessed the widespread of portable equipment from cellular phones to multimedia portable terminals. However, these mobile equipment are constrained in computational capability due to battery limitations and size limitations [53].

Over the last 30 years battery capacity has increased by a factor of 2 to 4, while the computational power of digital IC's increased more than 4 orders of magnitude [56]. The energy density of the Ni Cd batteries used in portable terminals is about 20 Watt-Hour/Pound [57]. Battery capacity isn't expected to increase dramatically over the next few years. New battery technology such as Nickel-Metal-Hydride is expected to have a capacity of no more than 30-35 Watt-Hour/Pound.

With the increase in market demand for new capabilities and functionality in mobile equipment, new approaches are required to reduce the power dissipation and hence prevent the battery size from growing in tandem with computational complexity.

Until recently power consumption was not a high priority issue in the design of VLSI systems. Performance (speed) and cost (area) were the two metrics that governed the design of VLSI systems [58]. However, the need for longer battery life in future portable terminals has added power dissipation to the metrics that should be considered when designing a VLSI system.

Even though our main concern in this chapter is mobile applications, it should be noted that they are not the only factor driving the need for lower power dissipation. Power dissipation in cutting-edge immobile equipment has

reached a limit where any further increase in power dissipation will lead to significant increase in the cost of packaging and the cooling system. The addition of a heat sink could increase the component cost by \$5-\$10 [59]. In addition, large power dissipation leads to lower component reliability. Every 10°C increase in temperature doubles the component failure rate [56].

Finally, there are the economical and environmental advantages of reducing the power dissipation. A study in 1993, [60] showed that the 60 million personal computers in the USA dissipated \$2 billion of electricity per year, and that they indirectly produced as much CO<sub>2</sub> as 5 million cars. In 1993 personal computers accounted for 5% of the commercial electricity demand, this is expected to increase to 10% by the year 2000.

Reducing the power dissipation can be done at the various levels of the design process, starting at the algorithmic and architectural levels and going down to the circuit and device levels [59]. The power minimization problem at each one of these levels has different characteristics and meets different challenges. At the higher design levels, the designer faces alternative choices with little information about the design parameters of the lower layers. At the lower design levels, the number of parameters is limited making the low-power design problem easier.

However, implementation of the low-level low-power design techniques requires greater investment and longer time to implement than the high-level low-power design techniques. Consider, for example, process scaling as a technique to reduce power dissipation. This requires a greater investment and a longer time to implement than changing the algorithm as a means of reducing the power dissipation.

Despite their great potential for reducing the power dissipation, high-level techniques are the least investigated techniques. Selecting the suitable algorithm and mapping it to the appropriate architecture can have a great influence on the minimization of power dissipation [61] [62]. Eliminating redundant and irrelevant computations has a substantial effect on the reduction of the power dissipation.

Future portable terminals are required to handle multimedia information - speech, video and data [62]. Because of the limited bandwidth allocated to mobile systems, compression/decompression of information is required in mobile terminals. Compression algorithms and in particular video compression algorithms demand large computation capability [63] which in turn leads to high power dissipation. The desire to have multimedia portable equipment has motivated work towards low-power implementations of video compression algorithms.

The organization of this chapter is as follows, section 9.2 talks about the sources of power dissipation in CMOS circuits and the parameters they depend on. In section 9.3, some of the techniques used in the estimation of the

power dissipation are presented. In section 9.4, we present some low-power design techniques developed for wireless portable systems and digital signal processors. Low-power techniques, used at the device and circuit levels, are presented in section 9.5.

The quadratic dependency of the power dissipation on the voltage makes voltage reduction an effective way to reduce the power dissipation. This is examined in section 9.6. However, reducing the voltage leads to longer delays. Techniques used to maintain a constant throughput with voltage scaling are also considered in section 9.6. Low-power techniques used at the architecture and algorithmic levels are presented in section 9.7.

## 9.2 SOURCES OF POWER DISSIPATION

The power dissipated in an electronic system depends on the implementation technology and the circuit style used. Current mode BJT and NMOS have DC (static) power dissipation, while CMOS has very low DC power dissipation, making its power dissipation lower than the two former technologies. The CMOS style is the most commonly used style for the implementation of VLSI systems.

In digital CMOS circuits, there are three sources of power dissipation [57]:

1. Switching power dissipation.
2. Short-circuit-current power dissipation.
3. Leakage-current power dissipation.

The most dominate of these is the switching power dissipation which is given by [64]:

$$P_{switching} = \alpha_{0 \rightarrow 1} C_L V_{DD}^2 f_{clk} \quad (9.1)$$

Where,

$\alpha_{0 \rightarrow 1}$  is the switching activity factor,

$C_L$  is the load capacitance,

$V_{DD}$  is the supply voltage,

$f_{clk}$  is the clock frequency.

The **switching power dissipation**, as seen from the previous equation, depends on four parameters. The switching activity factor  $\alpha_{0 \rightarrow 1}$ , the load capacitance  $C_L$ , the supply voltage  $V_{DD}$  and the clock frequency  $f_{clk}$ . Of these, the supply voltage has the greatest effect on the switching power dissipation

because of the quadratic dependence. In a well designed CMOS circuit, the switching power dissipation accounts for 90% of the power dissipation.

The switching activity factor  $\alpha_{0 \rightarrow 1}$ , the probability of a zero-one transition, depends on:

1. The logic function. For example, a NAND gate with equi-probable and independent inputs has a switching activity factor at its output given by:

$$\alpha_{0 \rightarrow 1} = \frac{3}{16},$$

while an XOR gate, with equi-probable and independent inputs has a switching activity factor at its output given by:

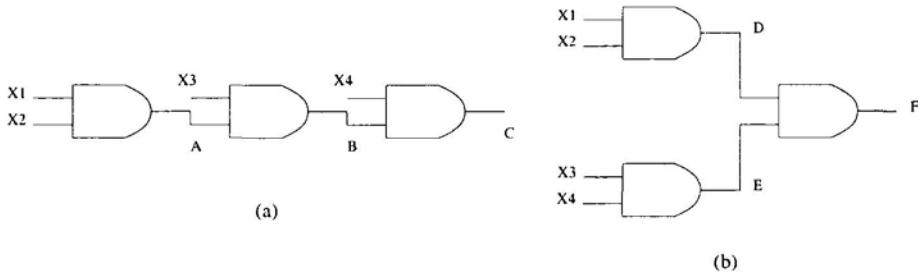
$$\alpha_{0 \rightarrow 1} = \frac{1}{4}.$$

2. The logic style. Dynamic logic has higher switching activity than static logic, because the output is precharged at the end of each cycle. However, dynamic logic is glitch free. The logic style also influences the capacitance.
3. The signal statistics. The higher the correlation between the successive samples the lower the switching activity.
4. The circuit topology. e.g. chain structure versus tree structure. A Chain structure has lower switching activity, but higher glitching power.

To see how the switching activity of a chain structure is lower than that of a tree structure, let's consider the case of building a four-input AND gate using two-input AND gates. The chain and the tree topologies are shown in Figure 9.1. For the chain structure, the output node of each AND gate, A, B and C, has a switching activity factor ( $\alpha_{0 \rightarrow 1}$ ) given by  $\frac{3}{16}$ ,  $\frac{7}{64}$ , and  $\frac{15}{256}$  respectively. While for the tree structure, the output node of each AND gate, D, E and F, has a switching activity factor ( $\alpha_{0 \rightarrow 1}$ ) given by  $\frac{3}{16}$ ,  $\frac{3}{16}$ , and  $\frac{15}{256}$  respectively. Notice that, the tree structure has a higher switching activity factor because node E has a higher switching activity than node B. However, for the tree structure, the delays to the input of each AND gate are equal. This is not true for the chain structure. Hence, the chain structure is susceptible to glitches, which increases the switching activity and hence the power dissipation.

The **short-circuit-current power dissipation**, unlike the switching power dissipation, depends on the rise and fall times of the input signal. To minimize the effect of the short-circuit power dissipation it is desirable to have equal input and output edge times [65]. In this case, the power dissipation is less than 10% of the total dynamic power dissipation.

The **leakage-current power dissipation** is due to:



*Figure 9.1.* Four-input AND gate built using two-input AND gates. (a) Chain Structure. (b) Tree Structure.

1. Reverse-bias diode leakage current. This is in the order of  $25 \mu A$  for a 1 million transistor chip. Hence, it represents a negligible component of the power dissipation [64].
2. Subthreshold current. Associated with this is the subthreshold slope  $S_{th}$ , which is the voltage required to reduce the subthreshold current by an order of magnitude [66]. The absolute minimum of  $S_{th}$  is about  $60 \text{ mv}/(\text{decade current})$  at room temperature. This can be achieved by using Silicon-On-Insulator (SOI) technology [67]. Lowering the subthreshold voltage increases this component.

Of these three power dissipation components, the switching power dissipation component is the most dominant [64]. Hence, this is the component we usually seek to minimize in digital CMOS circuits, especially at the architectural and algorithmic levels.

### 9.3 ESTIMATING THE POWER DISSIPATION

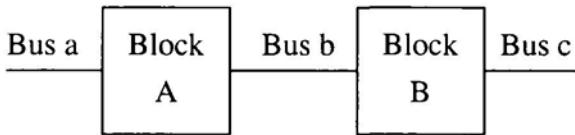
Power estimation can be a complex task. Not only does it require knowledge about the technological parameters of the system under consideration such as the operating voltage, the physical capacitance, the circuit style, etc., but it also requires detailed knowledge of the signal statistics such as the data activity and the signal correlation.

The aim of power estimation is to find the average power dissipated in a system based on a certain model. Power estimation becomes more inaccurate as the degree of model abstraction increases. Hence, the most accurate power estimators are the circuit simulators [68]. However, circuit simulators are slow and require complete and specific information about the inputs [69].

Gate-level probabilistic techniques have been proposed, ranging from simple techniques [70] which assume a zero-delay gate model and thus don't calculate the glitching power which can be as high as 70% [71], to more elaborate

techniques [72] [73] that not only consider the effect of glitching, but they also take into account the effect of temporal and spatial correlation [69].

The research done on power estimation at the higher abstraction levels is still limited [74] [75]. At the architecture level, the system is described in terms of interconnected operators (adders, multipliers, etc.) and memory blocks (registers, ROMs, etc.). These building blocks, as they will be called from now on, are interconnected by busses. Figure 9.2, shows a simplified architecture consisting of: two building blocks, one interconnecting bus (bus b), one input bus (bus a), and one output bus (bus c).



*Figure 9.2.* A simplified system consisting of two building blocks and the interconnection busses.

The total power dissipated in such an architecture is the sum of the power dissipated in the building blocks and the power dissipated in the busses. The power dissipated by a building block depends on:

1. Block activity factor ( $\beta$ ) (number of executions per second).
2. Output signal activity factor ( $\alpha$ ).
3. Normalized block energy  $E_n$ . The normalized energy is the energy dissipated by the building block per execution when the signal switching activity factor is one.

The total power dissipated by the building blocks is given by:

$$P_{BB} = \sum_{n \in \mathfrak{N}} \alpha_n \beta_n E_n \quad (9.2)$$

Where  $\mathfrak{N}$  is the set of all building blocks. The power dissipated by a bus depends on:

1. Signal activity factor ( $\alpha$ ).
2. Length of bus ( $\ell$ ).
3. Capacitance per unit length (C).

The total power dissipated by the busses is given by:

$$P_{BU} = K \sum_{n \in \mathfrak{N}} C_n \alpha_n \ell_n \quad (9.3)$$

Where  $\aleph$  is the set of all busses.  $K$  is some constant that depends on the operating voltage. When determining  $\ell_n$  the size of the building blocks should be taken into consideration.

## 9.4 LOW-POWER EXAMPLES OF PORTABLE SYSTEMS

There are numerous low-power techniques that have been developed by researchers and designers to lower the power dissipation of portable systems. Some of these techniques, developed for the design of wireless portable systems and digital signal processors, are presented in this section.

The first low voltage, very low current integrated circuits were developed about 25 years ago for the watch [76]. However, for other electronic systems power dissipation was only an afterthought. During the last decade, this has began to change. There has been great interest in the implementation of low-power, small size portable communicators for voice, video, images and data information as well as low-power note-book and lap-top computers [62] [77] [78] [79].

In cellular systems, a considerable fraction of battery energy is used for transmission. Reducing the cell size not only increases the spectrum efficiency through frequency reuse but it also allows operation at lower transmission power levels. This in turn leads to longer battery life. Currently mobile phones operate in a cell of several hundred meters radius, and transmit power in the order of 0.1-1 Watt [80].

The Viterbi decoder, used in CDMA cellular applications, presented in [81], employs various low-power techniques. The squared Euclidean measure has been substituted by a non-squared Euclidean measure. This reduces the complexity of the branch metric unit and the word-length of the path metric unit. The Viterbi decoder presented uses minimum sized processing units. To meet the throughput requirement, parallelism and pipelining are employed. To reduce spurious transitions on high-capacitance busses, gated control signals are used for controlling the multiplexers connected to these busses.

Surviving-path memory management [82] is one of the operations required in the Viterbi decoder. There are two techniques for surviving-path memory management: exchange register and trace back. In [83], the effect of hybrid techniques on reducing the power dissipation is considered.

In a receiver, the matched filter is positioned between the RF section and the baseband section. Hence, it can be implemented in digital or in analog technology. The effect of each implementation on the power dissipation is considered in [84]. It turns out that for slow matched filters, with a large number of taps and high precision, the digital implementation is more power efficient than the analog one.

By lowering the supply voltage from 5 volts to 1.5 volts, the power dissipation of different digital filters has been lowered by 8-11 times [85]. Architectural transformations such as parallelism, associativity, distributivity, commutativity, operation substitution and bit width optimization were used to maintain a constant throughput, lower the glitching activity, and reduce the interconnect capacitance.

In [86], a low-voltage low-power DSP is designed. The operating speed of the DSP is 63 MHz at 1 Volt. The power dissipation at this voltage and speed is 17.0 mW. During active operation of the DSP, power saving is realized by the use of locally gated clocks. Global gating is also available and it is controlled by three power-down instructions. The memory is divided into 8 arrays, only one array is activated during each memory access. A multi-level threshold voltage,  $V_T$ , is used. High  $V_T$  is used in the 6 transistors of the memory cells to lower the standby current. Low  $V_T$  is used in the peripheral circuitry to allow high-speed operation at 1 Volt.

A variable threshold voltage scheme is used in [87], to lower the standby power dissipation in a low  $V_T$  CMOS technology. It also mitigates the effect of fluctuations in  $V_T$  on the system delay. The threshold voltage is controlled by changing the substrate voltage  $V_{BB}$ . For the NMOS, in the active mode  $V_{BB} = -0.5$  Volt, and  $V_T = 0.1$  Volt. In the standby mode,  $V_{BB} = -3.3$  Volt and  $V_T = 0.5$  Volt.

## 9.5 REDUCING THE POWER DISSIPATION AT THE DEVICE AND CIRCUIT LEVELS

There are several techniques, during each level of the design process, to reduce the power dissipation. At the device level, the following techniques lead to lower power dissipation:

- **Silicon-On-Insulator (SOI) technology**, this leads to lower leakage currents and lower parasitic capacitance [88] [89].
- **Place and route optimization.** Assign signals with high switching activities to short wires. Also, assign global signals, such as the clock, to layers with low capacitance per unit length.
- **Transistor sizing.** Increasing (W/L) decreases the transistor delay which allows a decrease in voltage to maintain a constant throughput. However, increasing the transistor size increases the capacitance and hence the power dissipation. Hence, an optimum transistor size for minimum power dissipation exists. Assume that the interconnect capacitance is  $C_p$ , and the transistor input capacitance is  $C_i$ . It is found that, for small  $C_p/C_i$ , the optimum transistor size is the minimum size, otherwise there is an optimum size that gives minimum power dissipation [64].

- **Using submicron devices.** This reduces the parasitic capacitance and allows the use of a lower supply voltage, with minimum effect on the delay, for a velocity-saturated device [90].
- **Reducing the subthreshold voltage.** This allows a reduction in the operating voltage, with minimum effect on the delay. But this leads to larger subthreshold currents. Hence, a compromise is required. Some designs use a multi-threshold voltage technology [91]. While others use a variable threshold voltage [87].

At the circuit and logic levels, the following techniques can be used to reduce the power dissipation:

- **Reduce gate capacitance,** for example complementary pass-transistor logic has a lower input capacitance than conventional CMOS logic [92].
- **Reduced logic swing** [64] by making  $V_H = V_{DD} - V_T$ . However, this has two disadvantages:
  1. Low noise-margin-high ( $NM_H$ )
  2. Following gate can dissipate static power.
- **Low-power support circuitry**
  - Level converting circuit [64].
  - High efficiency low-voltage DC/DC converter [93].
- **Logic level power-down.** Modifying the circuits to allow power-down of unused logic blocks. This adds some overhead but can be beneficial if there are certain blocks that are not used for a large portion of the time [64] [94].
- **Multi-threshold circuit technology.** This allows the optimization of low-voltage circuits for high-speed and low-power [86] [95].
- **Scaled multi-buffer stages.** This compromises speed and power for gates driving large capacitive loads [57] [96].

## 9.6 LOW-VOLTAGE LOW-POWER OPERATION

The switching power is proportional to the square of the voltage, thus a quadratic reduction in power dissipation is achieved by lowering the supply voltage. However, the delay increases with the reduction of the supply voltage [53]. There are certain techniques used to keep the throughput constant despite the longer delay of the various building blocks. In this section, some of these techniques are investigated.

Figure 9.3.a shows the relative increase in delay as the supply voltage is scaled down. Figure 9.3.b shows the relative decrease in power dissipation as

the voltage is scaled down. Both figures were obtained for a CMOS inverter gate loaded by a 1 pF load and using  $0.8\mu\text{m}$  BiCMOS technology.

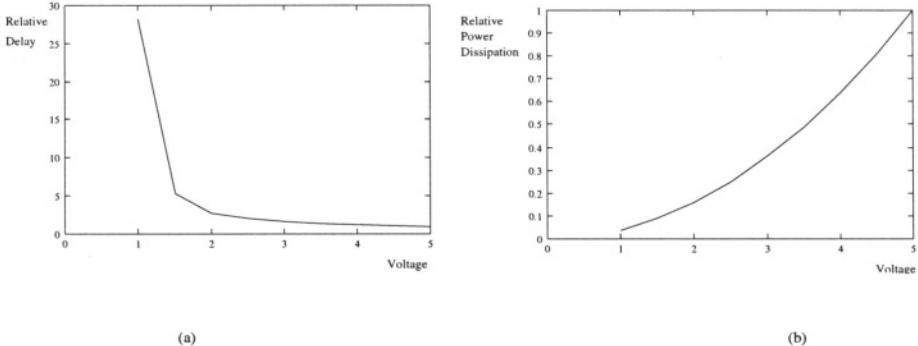
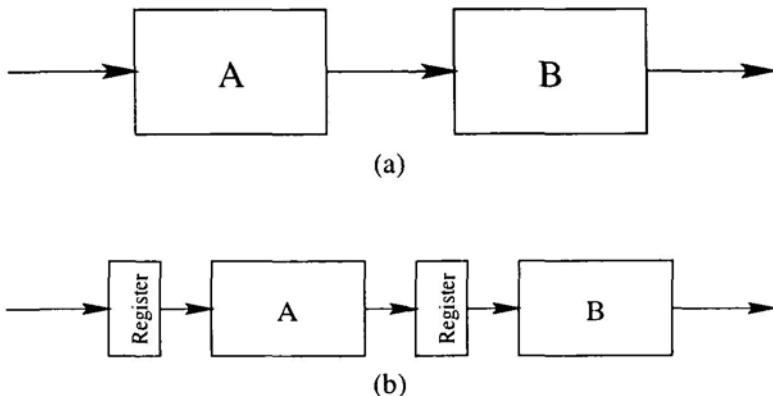


Figure 9.3. The effect of reducing the supply voltage in CMOS circuits, for a  $0.8\mu\text{m}$  BiCMOS technology. (a) Relative delay versus supply voltage. (b) Relative power dissipation versus supply voltage.

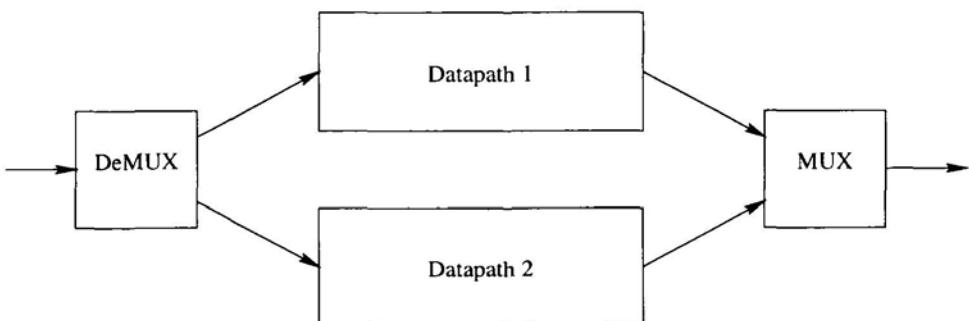
From Figure 9.3, it can be noticed that at low-voltage the rate of increase of the delay exceeds the rate of decrease of the power dissipation. This usually places a limit on the extent of using voltage scaling techniques. While lowering the supply voltage causes the delay to increase, performance can be regained by architectural modifications that speed up the system. Some of these architectural changes include, pipelining and parallelism [54], using faster adders, such as look-ahead adders or carry save adders [97] [98]. In this section, we consider the effect of using pipelining and parallelism with voltage scaling to lower the power dissipation while maintaining a constant throughput.

The architecture level, is the level in which operators (functional units) act on sets of logic values grouped into words. The manner in which these operators are interconnected or sequenced in time can have an influence on the performance of the architecture in terms of throughput, power dissipation and/or area, without affecting the actual functionality of the architecture.

For example, consider an architecture consisting of two cascaded operators as shown in Figure 9.4.a. Pipelining [54] this architecture, as shown in Figure 9.4.b, gives an alternative architecture with the same functionality as the original architecture, but with different performance. It is possible to operate the pipelined architecture at a lower supply voltage and at the same throughput as the original architecture. Thus through pipelining, it is possible to preserve the functionality and the throughput of the system but lower its power dissipation [53], at the cost of latency and extra overhead.



*Figure 9.4.* Two cascaded operators. (a) Non-pipelined architecture. (b) Two-stage pipelined architecture.



*Figure 9.5.* A two-datapath parallel system.

It is also possible, through parallelism, to decrease the power dissipation of the system [53]. In parallelism, the datapath is repeated  $N$  times, where  $N$  is the degree of parallelism. Like pipelining, the reduction of power dissipation in parallelism is due to the reduction in voltage, while the throughput is kept constant. Figure 9.5 illustrates parallelism.

The merits of pipelining over parallelism are:

1. Smaller area.
2. Reduced logic depth. Hence, less power due to glitches.

On the other hand, the disadvantage of pipelining over parallelism is the unbalanced pipe-stage delay problem [54]. To overcome this, we can combine parallelism and pipelining together as illustrated in the following example.

**Example: Combining Pipelining and Parallelism.** Consider two cascaded operators A and B as shown in Figure 9.4.a. Assume that:

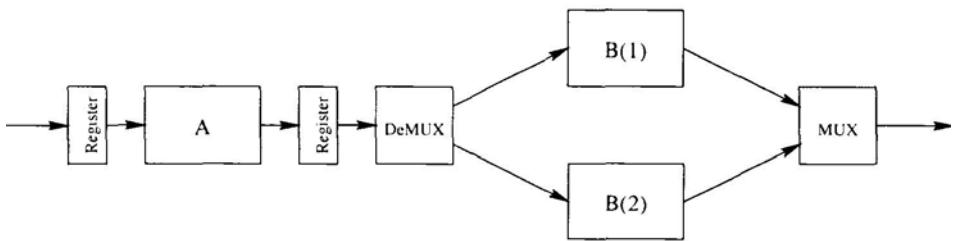


Figure 9.6. Combining parallelism with pipelining to balance pipe-stage delays.

$$D(A) = 1 \text{ unit}$$

$$D(B) = 2 \text{ units}$$

$$D(AB) = 3 \text{ units}$$

Where  $D(X)$  means the delay of operator  $X$ .

Pipeline this architecture as shown in Figure 9.4.b. Neglect the register delay with respect to that of the operator. The effective delay of the pipelined system is determined by the delay of the slowest stage which is 2 units in this case. It is possible now to reduce the voltage of the pipelined system to make the throughput of that system equal to the throughput of the non-pipelined one. The power dissipation, from Figure 9.3, is reduced by about 56%.

While the power was reduced by more than half of its value, yet we didn't make full use of pipelining due to the unbalanced pipe-stage delays. It is possible, by combining parallelism with pipelining, to balance the pipe-stage delays and hence achieve a larger reduction in power dissipation. Figure 9.6, shows a system that combines parallelism with pipelining. In this case, the effective system delay is 1 unit, allowing an 89% reduction in power dissipation, while maintaining the same throughput as the non-pipelined system.

There is a limit to pipelining and parallelism beyond which no improvement in power dissipation is possible, this is determined by:

1. The extra overhead required for pipelining and parallelism. This is represented by the pipeline registers for pipelining, and by the multiplexers, demultiplexers and the extra wiring capacitance for parallelism.
2. At low-voltage, the rate of increase of delay exceeds the rate of decrease of power dissipation.

The concept of parallelism can also be applied to memory accesses, where several bytes are accessed in parallel instead of accessing them sequentially. Parallelism in memory access is possible only if the data access pattern is sequential in nature [64].

## 9.7 REDUCING THE POWER DISSIPATION AT THE ARCHITECTURE AND ALGORITHM LEVELS

At the architecture and algorithm levels, the following techniques lead to lower power dissipation:

1. Reducing the switching activity, through proper choice of the number system [99]. For example, when the adjacent samples are highly correlated, the Gray code number system has a lower switching activity than the two's complement number system.
2. Using higher radix arithmetic. This reduces the number of block iterations per execution. This has a two-fold effect on the power dissipation. First, fewer block iterations means less power. However, there is a catch here, going to higher radix increases the block power per iteration, hence a compromise is required when choosing the power minimum radix. Second, fewer block iterations lead to higher throughput, which can be used to scale down the voltage to maintain a constant throughput while the power is reduced.
3. Using minimal sign-digit representation. This can be used to reduce the number of addition/subtraction operation per multiplication or division operation to reduce the power dissipation.
4. Using gated clocks and block deactivation.
5. Using suboptimal algorithms, that require substantially less computations per algorithm execution, while minimally degrading the performance of the algorithm. The remainder of this section is devoted to the study of suboptimal vector quantization decoding algorithms.

Vector quantization [44] is a compression technique. It exploits the correlation that exists between successive samples by quantizing a group of successive samples together. The encoder of a vector quantizer finds the representation vector closest to the quantized vector. The index of this vector is transmitted to the decoder. The decoder uses a look-up table or through computations finds the value of the representation vector.

The search process performed by the encoder, requires large computational capabilities to be performed exactly. However, approximate search algorithms exist, which greatly reduce the computational complexity with a much less degradation in performance. In this section, Full-Search Vector Quantization (FSVQ) and Tree-Structured Vector Quantization (TSVQ) [44] are considered.

Consider a vector  $X$  consisting of 8 samples. Each sample consists of 6 bits. This vector is to be approximated to the closest representation vector in set  $\{C_i\}$  of 64 representation vectors. That is, the compression ratio is 8:1.

**Table 9.1.** Computational complexity and memory requirement of VQ encoding algorithms. The VQ algorithm encodes a 64-level eight-sample vector into one of 64 representation vectors.

Algorithm	# of ADD/SUB	# of Mult.	# of Mem. access	ROM size
FSVQ	960	512	512	512 Byte
TSVQ $m = 2$	180	96	96	1008 Byte
TSVQ $m = 4$	180	96	96	672 Byte
TSVQ $m = 8$	240	128	128	576 Byte

The closest representation vector to vector X is the one having the least square error

$$E_i = \sum_{j=0}^7 (X_j - C_{ij})^2$$

Full-Search Vector Quantization, requires the calculation of the square error for every representation vector, then comparing the square errors to find the index of the representation vector with least square error. This always finds the nearest neighbor, however, it requires great computational capability as can be seen from Table 9.1.

In Tree-Structured Vector Quantization , the search is performed in stages [44]. During each stage, a subset of the representation vectors is eliminated from consideration by a relatively small number of operations. In general, consider a tree  $n$ -stages deep and having a branching factor  $m$  (the number of branches leaving a node) as shown in Figure 9.7. Each node in the tree has  $m$  vectors corresponding to each one of its  $m$  branches, the branch whose vector gives the least square error, is selected. The representation vectors corresponding to all the other branches are eliminated.

The total number of representation vectors is given by

$$N = m^n \quad (9.4)$$

The computational complexity of this algorithm is proportional to  $m \log_m(N)$  [44]. Minimum computational complexity is achieved at  $m = e$ . But  $m$  has to be an integer and preferably a power of two. Hence,  $m = 2$  or  $4$  is an optimum choice for minimum computational complexity. However, higher values of  $m$  give better performance but at the expense of greater computational complexity [44] [100]. Table 9.1 compares the computational requirements for various TSVQ algorithms and that of the FSVQ algorithm.

Notice from Table 9.1, that while the computational complexity decreases the ROM size increases which can lead to an increase in the power dissipation. Another factor in selecting the vector quantization algorithm is the performance of the algorithm, generally the lower the computational complexity, the lower the performance. However, for TSVQ the degradation in performance can be quite small [44]. In this case, a TSVQ decoder having  $m = 4$  represents

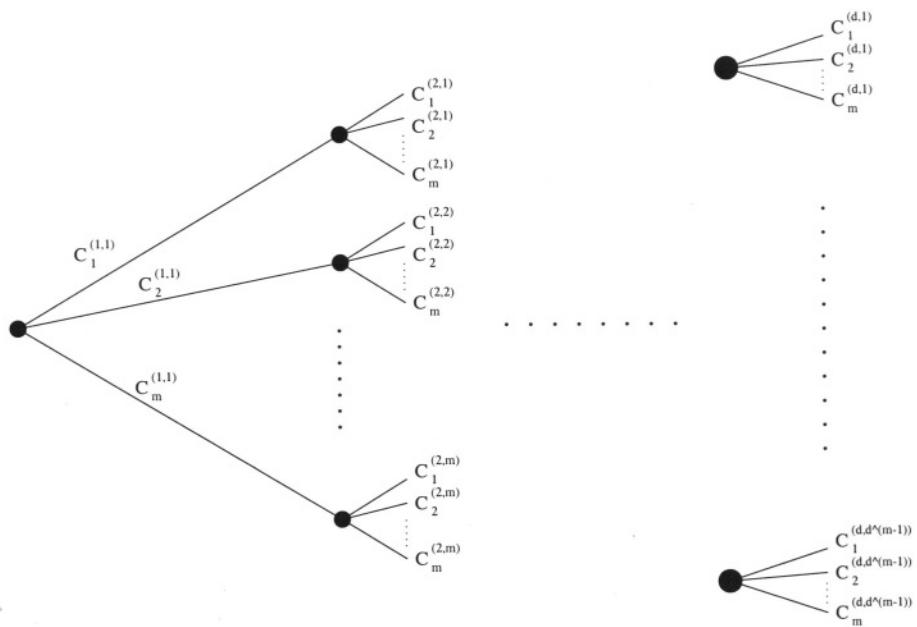


Figure 9.7. Tree-Structured Vector Quantization.

a good compromise between computational complexity (power dissipation), ROM size and algorithm performance.

Figure 9.7. Tree-Structured Vector Quantization.

a good compromise between computational complexity (power dissipation), ROM size and algorithm performance.

## Chapter 10

# AMPLIFIER DESIGN FOR WIRELESS COMMUNICATION SYSTEMS

### 10.1 INTRODUCTION

Amplification is a fundamental signal processing operation performed in any wireless communication system. Amplification is needed both in the transmitters and receivers. However, each amplifier has its own set of requirements and design criteria. In section 10.2, we talk about the general principles of amplifier design.

In the receiver, the amplifier is required to amplify a very weak signal that is buried in noise and interference, with the amplifier adding a minimum amount of noise so that the signal can be detected by the following signal processing stages. Low noise and amplifier gain are important parameters for determining the amplifier's performance. Low noise amplifiers are presented in section 10.3. Automatic gain control (AGC) amplifiers are also used in the receiver's amplification chain. These amplifiers are placed after the low noise amplifiers, and usually after down converting the received signal to an intermediate frequency. The purpose of the AGC amplifiers is to compensate for the variable attenuation of the wireless channel, so that the output of the amplifier has a constant root mean square (RMS) value. AGC amplifiers are presented in section 10.4.

On the other hand, in the transmitter, the signal is amplified to a sufficient level so that after it becomes attenuated, as it travels through the wireless channel, and noise and interference are added to it, the signal can still be detected at the receiver. Power amplification capabilities, power efficiency and linearity are important parameters for determining the amplifier's performance. Power amplifiers are presented in section 10.5

## 10.2 AMPLIFIER DESIGN

A single stage amplifier is an amplifier that consists of one active element. This active element may be a Bipolar Junction Transistor (BJT) or a Field Effect Transistor (FET). Each active element has three circuit configurations in which it can operate as an amplifier. In the common emitter configuration, the emitter of the BJT is connected to the ground, the input signal is applied to the base of the BJT, and the output signal is taken from the collector of the BJT. Figure 10.1.a shows the circuit diagram of a common emitter amplifier. This amplifier is an inverting amplifier that is capable of producing current and voltage gain.

In the common base configuration, the base of the BJT is connected to the ground, the input signal is applied to the emitter of the BJT, and the output signal is taken from the collector of the BJT. Figure 10.1.b shows the circuit diagram of a common base amplifier. This amplifier is a non-inverting amplifier capable of providing voltage gain only.

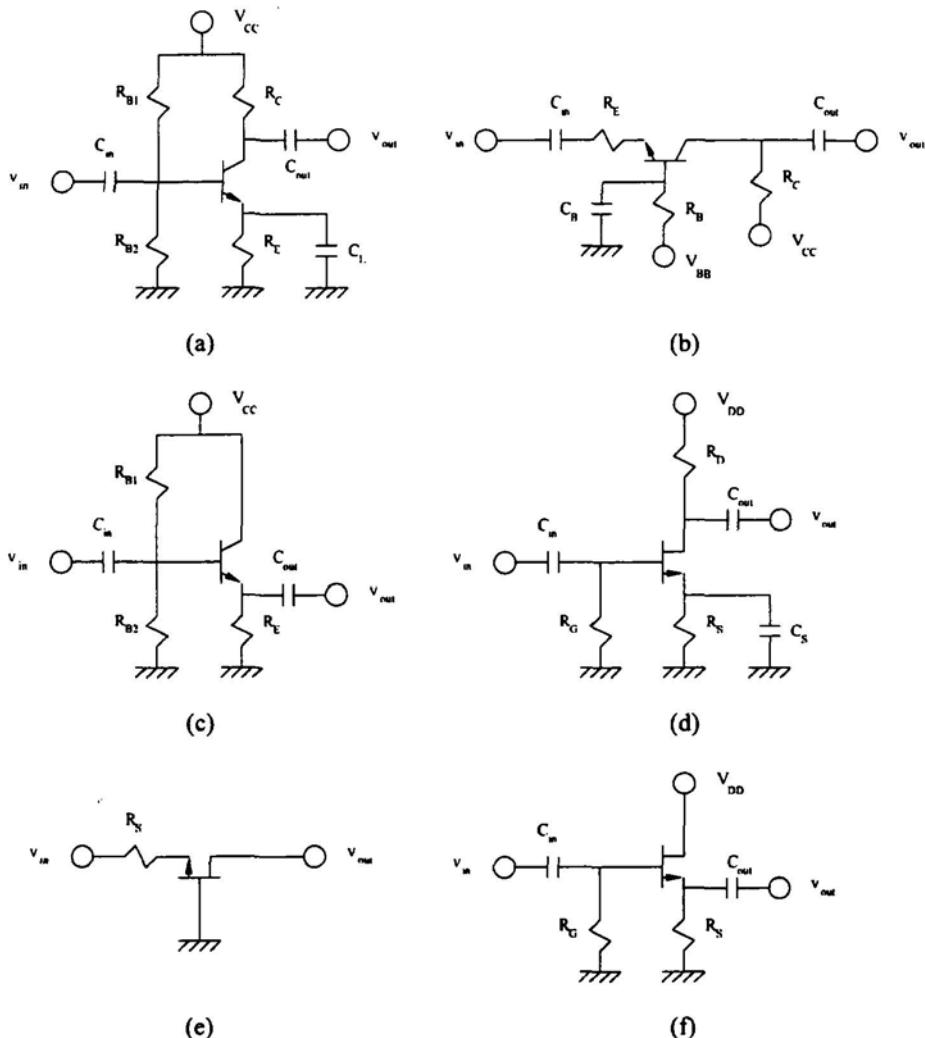
In the common collector configuration, the collector of the BJT is connected to the ground, the input signal is applied to the base of the BJT, and the output signal is taken from the emitter of the BJT. Figure 10.1.c shows the circuit diagram of a common collector amplifier. This amplifier, which is also known as an emitter follower, is characterized by its high input impedance. Hence, it is usually used as a buffer. This is a non-inverting amplifier that is capable of providing current gain only. This amplifier doesn't provide voltage gain.

In the common source configuration, the source of the FET is connected to the ground, the input signal is applied to the gate of the FET, and the output signal is taken from the drain of the FET. Figure 10.1.d shows the circuit diagram of a common source amplifier. This amplifier is an inverting amplifier that is capable of producing current and voltage gain.

In the common gate configuration, the gate of the FET is connected to the ground, the input signal is applied to the source of the FET, and the output signal is taken from the drain of the FET. Figure 10.1.e shows the circuit diagram of a common gate amplifier. This amplifier is a non-inverting amplifier capable of providing voltage gain only.

In the common drain configuration, the drain of the FET is connected to the ground, the input signal is applied to the gate of the FET, and the output signal is taken from the source of the FET. Figure 10.1.f shows the circuit diagram of a common drain amplifier. This amplifier, which is also known as a source follower, is characterized by its high input impedance. Hence, it is usually used as a buffer. This is a non-inverting amplifier that is capable of providing current gain only. This amplifier doesn't provide voltage gain.

To be able to design and analyze amplifier circuits, the transistor is replaced by a small signal model that models the transistor around its operating point.



**Figure 10.1.** Circuit diagrams of BJT and FET amplifiers. (a) Common Emitter. (b) Common Base. (c) Common Collector. (d) Common Source. (e) Common Gate. (f) Common Drain.

In this section, we present the high frequency models of the BJT and FET transistors. The high-frequency hybrid pi-model of the BJT transistor is shown in Figure 10.2.  $r_{bb'}$  is the base terminal resistance, this is usually in the range of 10 to  $50\Omega$ . This resistance is directly proportional to the width of the base. High frequency BJTs have a smaller base width than low frequency BJTs, hence  $r_{bb'}$  is smaller in high frequency transistors.  $r_{be'}$  is the base emitter junction resistance. This resistance is inversely proportional to the emitter DC current:

$$r_{b'e} = \frac{V_T \beta}{I_E} \quad (10.1)$$

Where,  $\beta$  is the forward current gain of the transistor. This is the ratio between the collector current and the base current.  $V_T$  is the thermal voltage. This is given by:

$$V_T = \frac{kT}{q} \quad (10.2)$$

Where,

$k$  is Boltzmann's constant,  $k = 1.38 \times 10^{-23}$  J/K.

$T$  is the temperature in Kelvin.  $T$  (Kelvin) =  $T$ (Celsius) + 273.16.

$q$  is the electron charge,  $q = 1.6 \times 10^{-19}$  C.

At room temperature (300 K),  $V_T$  is 26 mV. The base-collector junction capacitance  $C_{b'c}$  is formed due to the reverse bias across this junction. This capacitance depends on the base collector voltage as given by the following equation:

$$C_{b'c} \propto \frac{1}{V_{b'c}^{-n}} \quad (10.3)$$

Where  $n$  is between  $\frac{1}{2}$  to  $\frac{1}{3}$ . A diffusion capacitance exists across the base-emitter junction ( $C_{b'e}$ ). This capacitance increases linearly with the dc emitter current. Typically,  $C_{b'e}$  is much larger than  $C_{b'c}$ .

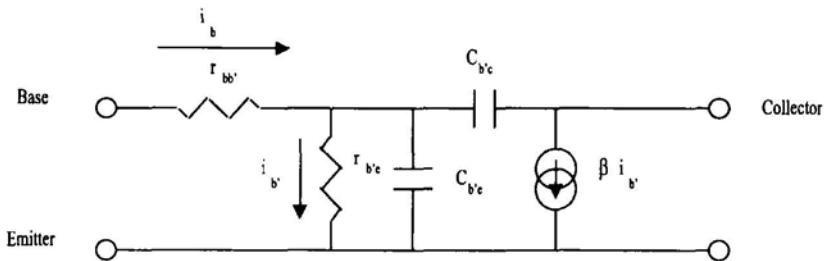


Figure 10.2. High-frequency hybrid pi-model of a bipolar junction transistor (BJT).

An important parameter that characterizes the high-frequency performance of the transistor is the short circuit unity gain frequency ( $f_T$ ). This is the frequency at which the (extrapolated) short circuit current gain of the transistor drops to unity. Consider short circuiting the emitter collector terminals in

Figure 10.2. Hence, the current  $i_c$  following through this short circuit and into the collector terminal is given by:

$$i_c = \beta i_{b'} - (jwC_{b'e})(i_{b'}r_{b'e}) \quad (10.4)$$

Where,

$$i_{b'}r_{b'e} = i_b \frac{r_{b'e}}{1 + jwr_{b'e}(C_{b'e} + C_{b'c})} \quad (10.5)$$

Therefore,

$$\begin{aligned} \beta(w) &= \frac{i_c}{i_b} \\ &= \frac{\beta - jwr_{b'e}C_{b'c}}{1 + jwr_{b'e}(C_{b'e} + C_{b'c})} \end{aligned} \quad (10.6)$$

The last equation indicates the presence of a zero and a pole in the transfer function of the short circuit current gain. However, the frequency of the zero is much larger than the frequency of the pole and hence its effect can be neglected in the frequency band of interest. The angular frequency of the pole is given by:

$$w_\beta = \frac{1}{r_{b'e}(C_{b'e} + C_{b'c})} \quad (10.7)$$

$w_\beta$  is the frequency at which the short circuit current gain is 3 dB lower than its low frequency value, assuming that the frequency of the zero is much larger than that of the pole. The short circuit current gain can be written as:

$$\beta(w) = \frac{\beta}{1 + jw/w_\beta} \quad (10.8)$$

Assume that  $w_T$  is the angular frequency at which the short circuit current gain is equal to unity. Hence,  $w_T$  and  $w_\beta$  are related by the following equation:

$$w_T = \sqrt{\beta^2 - 1}w_\beta \approx \beta w_\beta \quad (10.9)$$

Using equation 10.7, the relation between the short circuit unity gain frequency ( $f_T$ ) and the device parameters of a BJT transistor is given by:

$$f_T = \frac{\beta}{2\pi r_{b'e}(C_{b'e} + C_{b'c})} \approx \frac{\beta}{2\pi r_{b'e}C_{b'e}} \quad (10.10)$$

The last approximation is valid because typically:  $C_{b'e} \gg C_{b'c}$ .

A similar high frequency hybrid pi-model exists for the field-effect transistor (FET). There are two types of FETs, these are the junction field-effect transistor (JFET) and the metal-oxide semiconductor field-effect transistor (MOSFET). To operate as an amplifier, the FET is biased so that it is in saturation. In saturation, the drain current  $I_d$  of the FET depends primarily on the gate-source voltage ( $V_{gs}$ ) and to a much lesser extent on the gate-drain voltage ( $V_{ds}$ ). The drain current, in saturation, is given by:

$$I_d = K(1 + \lambda V_{ds})(V_{gs} - V_{th})^2 \quad (10.11)$$

Where,

$V_{th}$  is the threshold voltage of the FET.

$\lambda$  is the channel-length modulation parameter [101].

$k$  is a constant that depends of the characteristics of the transistor.

Ideally,  $\lambda$  should be zero, and the drain current flowing during saturation becomes independent of  $V_{ds}$ . However, in reality  $\lambda$  ranges between  $0.01V^{-1}$  to  $0.1V^{-1}$ . The square-law relation given by equation 10.11 accurately models the characteristics of a MOSFET transistor operating in saturation. A JFET transistor is more accurately modeled by a  $\frac{3}{2}$ -power relation, however equation 10.11 is commonly used as an approximation for the characteristics of a JFET operating in saturation [102].

Figure 10.3 shows the hybrid pi-model of a field-effect transistor.  $C_{gs}$  and  $C_{gd}$  are the gate-source and gate-drain capacitance respectively. Both these capacitors arise due to a reverse biased PN junction in the case of a JFET transistor, and due to the oxide capacitance in the case of a MOSFET transistor. Typically,  $C_{gs}$  is much larger than  $C_{gd}$ . In addition to these two capacitance, there also exists parasitic capacitance between the substrate and each terminal of the transistor (the source, gate and drain). The transconductance ( $g_m$ ) of the FET is given by:

$$g_m = \frac{2I_d}{V_{gs} - V_{th}} \quad (10.12)$$

Notice that, the transconductance depends on the drain current as well as the gate-source voltage. The output resistance of the FET depends on the drain current and is given by:

$$r_o = \frac{1}{\lambda I_d} \quad (10.13)$$

Typically,  $r_o$  is in the range of  $100K\Omega$ . The unity current gain frequency of the field-effect transistor is given by:

$$f_T = \frac{1}{2\pi} \frac{g_m}{C_{gs} + C_{gd}} \quad (10.14)$$

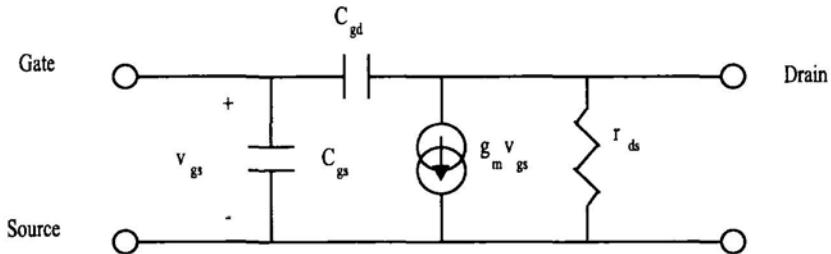


Figure 10.3. High-frequency hybrid pi-model of a field effect transistor (FET).

### 10.3 LOW NOISE AMPLIFIER

The noise level sets the ultimate limit of detectability of weak signals. This noise can be due to fluctuations in the received signal around the desired value because of thermal noise or other types of radio frequency interference. This noise can also be due to noise added by the signal processing operations performed in the receiver. Hence, it is desirable to use low noise signal processing components.

The first signal processing stage in the receiver after the duplexer is the low noise amplifier (LNA). The LNA needs to be designed to minimize the amount of noise added to the signal, i.e. have a low noise figure, and to have a high gain such that the signal is amplified to a level that would make the noise of the subsequent signal processing stages negligible.

Thermal noise, also known as Johnson noise, is generated across the terminals of any resistor. Thermal noise has a flat spectrum, up to a certain frequency. Hence, the noise voltage generated across the resistor increases as the bandwidth increases. In addition to having a flat spectrum (white noise), thermal noise also has a Gaussian distribution. The RMS thermal noise voltage generated across the terminals of a resistor is given by:

$$V_n = \sqrt{4kTRB} \quad (10.15)$$

Where,

$k$  is Boltzmann's constant,  $k = 1.38 \times 10^{-23}$  J/K.

$T$  is the temperature in Kelvin.

$R$  is the resistance of the resistor.

$B$  is the bandwidth.

Another source of noise in electronic circuits is shot noise. Shot noise occurs when the carriers change energy states as they move across a PN junction. Because the charges that carry the current are finite, any statistically variations in the number of charges that carry the current per unit time causes randomness in the current which is known as shot noise. The RMS shot noise current generated due to a DC current  $I_{DC}$  is given by:

$$I_n = \sqrt{2qI_{DC}B} \quad (10.16)$$

Where,

$q$  is the electron charge,  $q = 1.6 \times 10^{-19}$  C.

$B$  is the bandwidth.

Notice that, the ratio between the RMS value of the noise current to the DC current increases as the value of the DC current decreases. For example assuming a 1 MHz bandwidth. The RMS shot noise current for a DC current of 1 nA is 40 pA, which is 4% of the DC current. However, the RMS value of the shot noise for a DC current of 1 pA is 0.566 pA, which is 56.6% of the DC current. Shot noise, like thermal noise, is both white and Gaussian.

Another source of noise in electronic circuits is the flicker noise. Flicker noise has a  $1/f$  spectrum, and is sometimes called pink noise.

The noise figure, which is used to characterize the performance of an amplifier, is defined as the ratio, in decibels, of the output of the actual (noisy) amplifier, to that of an ideal noiseless amplifier having the same gain, and with the same resistor  $R_s$  connected across the input terminals of each amplifier.

Figure 10.4.a shows a noisy two-port network, this network can be modeled by a noise-less network having two noise sources connected to its input.  $e_n$  represents the voltage noise source. This noise source is connected in series with the input terminal.  $i_n$  represents the current noise source. This noise source is connected in parallel to the input terminal.

The noise current source generates a noise voltage across the source resistance that is given by:  $i_n R_s$ . Assuming that the two noise sources are uncorrelated, hence we can add the square of each to produce the squared effective noise voltage:

$$e_T^2 = e_n^2 + (i_n R_s)^2 \quad (10.17)$$

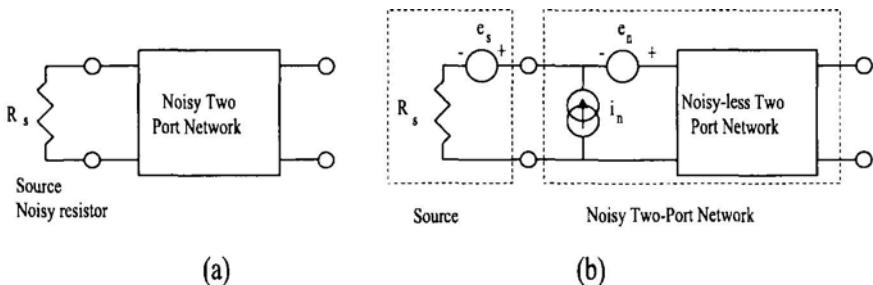


Figure 10.4. (a) A noisy two-port network. (b) The equivalent two-port network, having the noise sources lumped in  $v_n^2$  and  $i_n^2$ .

The source resistance generates a noise voltage whose squared value is given by:

$$e_s^2 = 4kTR_sB \quad (10.18)$$

Hence, the noise figure of the two-port network shown in Figure 10.4 can be expressed as:

$$NF = 1 + \frac{e_n^2 + (i_n R_s)^2}{4kTR_sB} \quad (10.19)$$

Equation 10.19 indicates that when the source resistance is low, the voltage noise source is the dominant contributor to the noise figure. In this case, it is desirable to use a two-port network having a lower  $e_n$ . On the other hand, if  $R_s$  is high, the current noise source is the dominate contributor to the noise figure. In this case, it is desirable to use a two-port network having a lower  $i_n$ . The optimum  $R_s$  at which the noise figure is minimum is given by:

$$R_{s,(opt)}^2 = \frac{e_n^2}{i_n^2} \quad (10.20)$$

Practically, it is unlikely to have the source resistance equal to the optimum resistance. Typically, the source resistance is  $50\Omega$  to  $75\Omega$ . The optimum resistance is much higher than this. Hence, a matching network is used to transform the source resistance to be equal to the optimum noise figure resistance.

The bipolar junction transistor can be modeled by a noise-less transistor with an equivalent voltage noise source and an equivalent current noise source

connected to its input. The equivalent voltage and current noise sources are typically in the range of nano-volts and pico-amps per  $\text{Hz}^{1/2}$  respectively.

There are two sources that contribute to the equivalent voltage noise source of the bipolar junction transistor. First, the thermal noise of the base resistance  $r_{bb'}$ . Second, the shot noise across the base-emitter junction. These two noise sources are uncorrelated, hence, their noise powers (square of the noise voltage) can be added. Accordingly, the voltage noise source of the BJT is given by:

$$\begin{aligned} e_n^2 &= 4kT r_{bb'} B + 2qI_c r_e^2 B \\ &= 4kT r_{bb'} B + \frac{2(kT)^2}{qI_c} B \end{aligned} \quad (10.21)$$

According to the last equation, the noise voltage of a bipolar junction transistor decreases with the increase of the DC collector current, because of the decrease in shot noise. This is true to a certain limit, beyond that the flicker noise, arising from the base resistance  $r_{bb'}$  becomes a dominate noise source. This noise source increases with the increase of the base current, and accordingly the collector current. The noise voltage decreases slightly with the increase of frequency because of the decrease of flicker noise with frequency. This effect is more apparent at high collector currents where the flicker noise is more dominant.

The main contributors to the current noise source is shot noise, and flicker noise. Neglecting the effect of flicker noise, the noise current is given by:

$$i_n^2 = 2qI_b B \quad (10.22)$$

Accordingly, the noise current increases with the increase of the base (and collector) DC currents. As the frequency decreases, the flicker noise becomes dominate, because of its  $1/f$  spectrum, and according the noise current starts increasing.

Knowing that,  $r_e = 1/g_m$  and  $V_T = kT/q$ , it can be shown that, the optimum source resistance to minimize the noise figure, according to equation 10.20 is given by:

$$R_{s,\text{opt}} = \frac{\beta^{1/2}}{g_m} \sqrt{1 + 2g_m r_{bb'}} \quad (10.23)$$

Notice that, because  $g_m$  is directly proportional to the collector current, the optimum source resistance increases with the increase of collector current. The optimum (minimum) noise figure is given by:

$$\text{NF}_{\min} = 1 + \frac{1}{\beta^{1/2}} \sqrt{1 + 2g_m r_{bb'}} \quad (10.24)$$

Since, the noise voltage drops as the collector current increases, while the noise current increases as the collector current increases. It is possible to optimize the transistor's operating current such that the total noise introduced by the transistor is minimized for a particular source resistance.

The field effect transistor can also be modeled as a noise-less transistor with an equivalent voltage noise source and an equivalent current noise source connected to its input. There are two sources that contribute to the equivalent voltage noise source of the field effect transistor. First, the thermal noise of the channel resistance. Second the flicker noise due to surface traps. This noise has a  $1/f$  spectral dependency. These two noise sources are uncorrelated, hence, their noise powers (square of the noise voltage) can be added. Accordingly, the equivalent voltage noise source of a FET is given by:

$$e_n^2 = \frac{8kT}{3g_m} B + \frac{K_f I_d^n}{f} B \quad (10.25)$$

Where,

$g_m$  is the FET's transconductance.

$K_f$  is a proportionality constant for the flicker noise.

$a$  is an exponential factor.  $0.5 < a < 2$ .

As the drain current increases the FET's transconductance increases, accordingly the channel's thermal noise voltage decreases. Furthermore, as the width-to-length ratio ( $W/L$ ) of the FET transistor increases, the channel's thermal noise decreases. The flicker noise decreases as the product  $WL$  increases.

The main contributor to the current noise source is shot noise due to the leakage current. This is given by:

$$i_n^2 = 2qI_g \quad (10.26)$$

Where,  $I_g$  is the gate leakage current. Usually, this current is very small and the FET's noise current can be neglected. This makes the FET achieve good noise performance for high source resistances. Since, the gate leakage current increases with temperature, the noise current  $i_n$  also increases with temperature.

Bipolar junction transistors have a lower noise voltage than field effect transistors. This makes BJTs have a superior noise performance with low source resistance. However, as the source resistance becomes higher, and the noise current becomes the dominate source of noise added by the transistors, FETs start having superior noise performance.

## 10.4 AUTOMATIC GAIN CONTROL AMPLIFIERS

In a wireless communications system, the location of the mobile terminal relative to the base station is always changing. Since, the received power level changes with the nth power of the distance (n usually ranges between 2 to 6 as shown in Table 3.1), hence there can be wide variations in the received power level. This necessitates the use of automatic gain control (AGC) amplifiers, to limit the dynamic range requirements of subsequent signal processing stages.

The automatic gain control is achieved by using an amplifier whose gain changes as the received signal strength changes in such a way so as to make output signal strength of the AGC amplifier constant. A DC control voltage determines the gain of the AGC amplifier. This DC control voltage is obtained from the receivers output signal.

The differential amplifier shown in Figure 10.5, is an example of an AGC amplifier. The AGC control voltage is applied to the base of the constant current source transistor ( $Q_3$ ). While, the differential input signal is applied to the bases of transistors  $Q_1$  and  $Q_2$ . As the AGC voltage increases, the current of the constant current source ( $I_{EE}$ ) increases. This decreases the emitter resistance of  $Q_1$  and  $Q_2$ , and increases the gain of the differential amplifier.

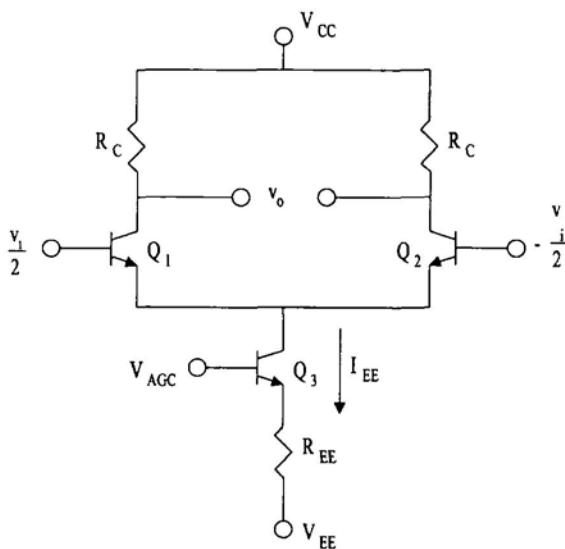


Figure 10.5. A gain-controlled differential amplifier.

The output signal of the amplifier shown in Figure 10.5 is a non-linear function of the input signal. To linearize this amplifier an emitter resistance

$R_E$  is inserted as shown in Figure 10.6. This emitter resistance, introduces local negative feedback, which linearize the differential amplifier.

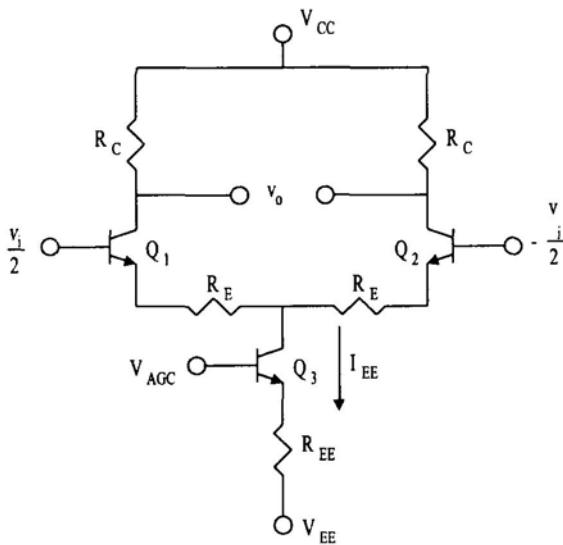


Figure 10.6. A gain-controlled linearized differential amplifier.

Assume that the input to the amplifier is a differential signal, such that the voltage at the base of  $Q_1$  is  $v_i/2$ , and that at the base of  $Q_2$  is  $-v_i/2$ . Furthermore, assume that the emitter current of  $Q_1$  is  $I_{E1}$  and the emitter current of  $Q_2$  is  $I_{E2}$ . The base emitter voltage drop ( $V_{BE}$ ) and the emitter current ( $I_E$ ) for any bipolar junction transistor are related by the following exponential expression:

$$\begin{aligned} I_E &= I_{ESS} \left( e^{V_{BE}/V_T} - 1 \right) \\ &\approx I_{ESS} e^{V_{BE}/V_T} \end{aligned} \quad (10.27)$$

Where,

$V_T$  is the thermal voltage (approximately 26 mV at room temperature).

$I_{ESS}$  is the emitter saturation current.

Hence, for the circuit shown in Figure 10.6:

$$v_i = V_T \ln \left( \frac{I_{E1}}{I_{E2}} \right) + (I_{E1} - I_{E2}) R_E \quad (10.28)$$

Knowing that:

$$I_{EE} = I_{E1} + I_{E2} \quad (10.29)$$

and (assuming that the base currents are negligible)

$$v_o \approx (I_{E1} - I_{E2})R_c \quad (10.30)$$

Hence the relationship between  $v_o$  and  $v_i$ , for the circuit shown in Figure 10.6, is given by:

$$v_i R_c = V_T R_c \ln \left[ \frac{I_{EE} R_c + v_o}{I_{EE} R_c - v_o} \right] + v_o R_E \quad (10.31)$$

Equation 10.31 indicates that the relationship between  $v_i$  and  $v_o$  is a non-linear one.  $R_E$  helps to reduce this non-linearity. To show this, we are going to determine the second and third order intermodulation distortion as a function of  $R_E$ . Differentiating equation 10.31 with respect to  $v_i$ , we get:

$$R_c = \frac{V_T R_C}{I_{EE} R_C + v_o} \frac{dv_o}{dv_i} + \frac{V_T R_C}{I_{EE} R_C - v_o} \frac{dv_o}{dv_i} + R_E \frac{dv_o}{dv_i} \quad (10.32)$$

Therefore, at  $v_o = 0$  (the operating point of the circuit):

$$\frac{dv_o}{dv_i} = \frac{R_C I_{EE}}{2V_T + R_E I_{EE}} \quad (10.33)$$

Differentiating equation 10.32 with respect to  $v_i$ , we get the second derivative of  $v_o$  with respect to  $v_i$ :

$$\begin{aligned} 0 &= \frac{V_T R_C}{I_{EE} R_C + v_o} \frac{d^2 v_o}{dv_i^2} - \frac{V_T R_C}{(I_{EE} R_C + v_o)^2} \left( \frac{dv_o}{dv_i} \right)^2 \\ &+ \frac{V_T R_C}{I_{EE} R_C - v_o} \frac{d^2 v_o}{dv_i^2} + \frac{V_T R_C}{(I_{EE} R_C - v_o)^2} \left( \frac{dv_o}{dv_i} \right)^2 \\ &+ R_E \frac{d^2 v_o}{dv_i^2} \end{aligned} \quad (10.34)$$

Therefore, at  $v_o = 0$ :

$$\frac{d^2 v_o}{dv_i^2} = 0 \quad (10.35)$$

Differentiating equation 10.34 with respect to  $v_i$ , we get the third derivative of  $v_o$  with respect to  $v_i$ :

$$\begin{aligned}
0 &= \frac{V_T R_C}{I_{EE} R_C + v_o} \frac{d^3 v_o}{d v_i^3} - \frac{V_T R_C}{(I_{EE} R_C + v_o)^2} \frac{d^2 v_o}{d v_i^2} \frac{d v_o}{d v_i} \\
&- \frac{2 V_T R_C}{(I_{EE} R_C + v_o)^2} \frac{d^2 v_o}{d v_i^2} \frac{d v_o}{d v_i} + \frac{2 V_T R_C}{(I_{EE} R_C + v_o)^3} \left( \frac{d v_o}{d v_i} \right)^3 \\
&+ \frac{V_T R_C}{I_{EE} R_C - v_o} \frac{d^3 v_o}{d v_i^3} + \frac{V_T R_C}{(I_{EE} R_C - v_o)^2} \frac{d^2 v_o}{d v_i^2} \frac{d v_o}{d v_i} \\
&+ \frac{2 V_T R_C}{(I_{EE} R_C - v_o)^2} \frac{d^2 v_o}{d v_i^2} \frac{d v_o}{d v_i} + \frac{2 V_T R_C}{(I_{EE} R_C - v_o)^3} \left( \frac{d v_o}{d v_i} \right)^3 \\
&+ R_E \frac{d^3 v_o}{d v_i^3}
\end{aligned} \tag{10.36}$$

Therefore, at  $v_o = 0$ :

$$\begin{aligned}
\frac{d^3 v_o}{d v_i^3} &= \frac{-4 V_T}{(I_{EE} R_C)^2} \frac{1}{2 V_T + I_{EE} R_C} \left( \frac{d v_o}{d v_i} \right)^3 \\
&= \frac{-4 V_T R_C I_{EE}}{(2 V_T + I_{EE} R_C)^4}
\end{aligned} \tag{10.37}$$

The transfer function of a non-linear amplifier (see equation 6.13 in chapter 6) can be expressed as:

$$v_{out} = a_1 v_{in} + a_2 v_{in}^2 + a_3 v_{in}^3 \tag{10.38}$$

Where:

$$\begin{aligned}
a_1 &= \frac{d v_o}{d v_i} \\
&= \frac{R_C I_{EE}}{2 V_T + R_E I_{EE}}
\end{aligned} \tag{10.39}$$

$a_1$  is the gain of the amplifier, and

$$\begin{aligned}
a_2 &= \frac{1}{2} \frac{d^2 v_o}{d v_i^2} \\
&= 0
\end{aligned} \tag{10.40}$$

and

$$a_3 = \frac{1}{6} \frac{d^3 v_o}{d v_i^3}$$

$$= \frac{-2V_T R_C I_{EE}}{3(2V_T + I_{EE} R_C)^4} \quad (10.41)$$

The third order intermodulation distortion is defined as the ratio between the third order non-linear component and the linear component. This is given by:

$$\begin{aligned} \text{IM}_3 &= \frac{a_3 v_i^3}{a_1 v_i} = \frac{a_3}{a_1^3} v_o^2 \\ &= -\frac{2V_T}{3(2V_T + R_E I_{EE})} \left( \frac{v_o}{I_{EE} R_E} \right)^2 \end{aligned} \quad (10.42)$$

Notice that, the presence of  $R_E$  reduces the intermodulation distortion of the amplifier. However, at the same time it also reduces the gain of the amplifier and increases its noise. This necessitates some type of compromise when selecting the value of  $R_E$ .

Figure 10.7 shows an alternative gain-controlled amplifier. In this circuit the input signal varies the emitter current  $I_{ee}$ . While the AGC voltage is applied to the base of transistor  $Q_1$ . This voltage steers the current between transistors  $Q_1$  and  $Q_2$ .

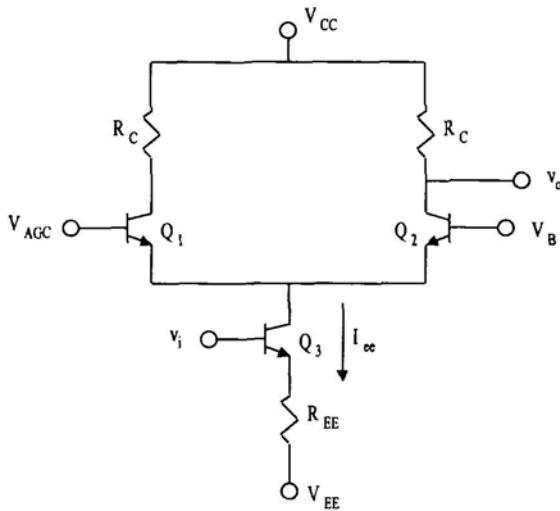


Figure 10.7. A gain-controlled amplifier. The input signal controls the emitter current.

To understand how this circuit works, let's begin by assuming that the output of the receiver is low. In this case,  $V_{AGC}$  is low, this steers most of the emitter current ( $I_{ee}$ ) through  $Q_2$ , making the gain of the amplifier high. On the

other hand, if the output of the receiver is high,  $V_{AGC}$  is high. This steers the emitter current away from the  $Q_2$ , making the gain of the amplifier low.

The current flowing through the transistor  $Q_3$  is given by:

$$I_{ee} = \frac{v_i - v_{be3} - V_{EE}}{R_{EE}} \quad (10.43)$$

To a first order approximation,  $I_{ee}$  is linearly dependant on the input voltage  $v_i$ . Strictly speaking however,  $v_{be3}$  depends exponentially on the emitter current of  $Q_3$ . Hence, the relationship between  $I_{ee}$  and  $v_i$  is non-linear. The emitter current flowing through transistor  $Q_2$  depends on the voltage difference,  $V_{AGC} - V_B$ :

$$I_{e2} = \frac{I_{ee}}{1 + e^{(V_{AGC} - V_B)/V_T}} \quad (10.44)$$

Neglecting the base current, the output voltage of the amplifier is given by:

$$v_o \approx V_{CC} - I_{e2}R_C \quad (10.45)$$

Combining equations 10.43, 10.44 and 10.45, we get the following expression relating  $v_o$  to  $v_i$ :

$$v_o = V_{CC} - \frac{(v_i - v_{be3} - V_{EE})R_C}{R_{EE}(1 + e^{(V_{AGC} - V_B)/V_T})} \quad (10.46)$$

The non-linearity in equation 10.46 is introduced due to the non-linear dependence of  $v_{be3}$  on  $v_i$ . However, this is a secondary non-linear effect that can be neglected. Hence, the circuit of Figure 10.7 is more linear than the circuit of Figure 10.5. Neglecting the voltage dependence of  $v_{be3}$  on  $v_i$ , the gain of the AGC amplifier of Figure 10.7 is given by:

$$\frac{dv_o}{dv_i} = -\frac{R_C}{R_{EE}(1 + e^{(V_{AGC} - V_B)/V_T})} \quad (10.47)$$

Notice that, as  $V_{AGC}$  increases, the AGC amplifier gain decreases and vice versa.

## 10.5 POWER AMPLIFIERS

The power amplifier is an essential building block in any wireless communication system. The power amplifier is the last signal processing block in the transmitter before the antenna. Only a band-pass filter and a duplexer (or a switch) follow the power amplifier, to suppress any out-of-band signals

generated by the power amplifier, and to combine the receive and transmit signals.

The power amplifier needs to amplify the RF signal to a high power level (hence its name), so that even after the signal is attenuated and corrupted by noise and interference it can be detected at the receiver. Because of this requirement (amplifying signals to a high power level), power amplifiers are the most power consuming block in a wireless communications system. Efficiency is an important parameter in determining how much power is consumed by the power amplifier to achieve a certain output power level. The efficiency of a power amplifier ( $\eta$ ) is defined as the ratio between the output RF power ( $P_{RF,out}$ ) to the power drawn from the DC power supply ( $P_{DC}$ ):

$$\eta = \frac{P_{RF,out}}{P_{DC}} \quad (10.48)$$

A more accurate expression for  $\eta$ , should also take the effect of the input RF power into account by adding it to  $P_{DC}$  in the denominator of equation 10.48. However, in most cases the input RF power is so small that its contribution can be neglected. Ideally,  $\eta$  should be unity, all the DC power supplied to the power amplifier is converted to RF power. However, practically,  $\eta$  is less than unity because a portion of the DC power is dissipated in the components used to build the power amplifier.

Another requirement imposed on power amplifiers is linearity. A power amplifier needs to be linear to avoid the generation of out-of-band harmonics. Linear power amplifiers have generally a lower efficiency than non-linear power amplifiers. There are different classes of power amplifiers that compromise between efficiency and linearity, in this section we look at some of these.

Class A power amplifiers are the simplest type of power amplifiers. They are designed so that the transistor always operates in the active (linear) region. Figure 10.8 shows an example of a class A power amplifier. The circuit is designed so that the DC operating point is in the middle of its linear region. The transistor is in saturation when  $v_o = v_{ce,sat}$ . The transistor is in cutoff when the collector current is zero, in this case  $v_o = V_{cc}$ . For the transistor to be in the middle of its linear region its DC operating point should be at:

$$v_{o,opt} = \frac{V_{cc} + v_{ce,sat}}{2} \quad (10.49)$$

Assume that an AC signal is injected into the power amplifier shown in Figure 10.8. The output of the amplifier consists of an AC signal superimposed on the DC voltage as shown in Figure 10.9. The maximum peak amplitude of the AC signal that the power amplifier can generate with the transistor being always in the linear region is:

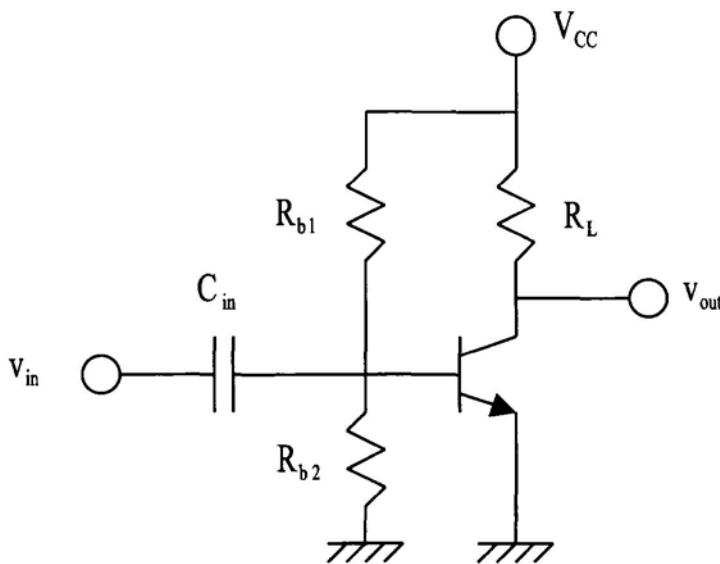


Figure 10.8. Class A power amplifier.

$$V_{AC,max} = \frac{V_{cc} - v_{ce,sat}}{2} \quad (10.50)$$

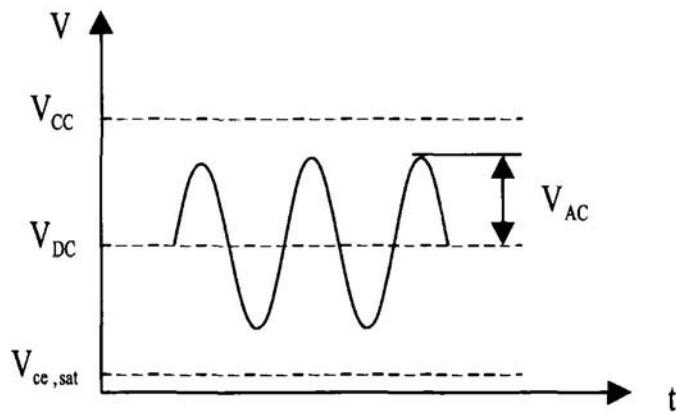


Figure 10.9. The output waveform of a class A power amplifier.

Hence, the maximum AC power delivered to the load resistor  $R_L$  is:

$$\begin{aligned}
 P_{AC,max} &= \frac{1}{2} \frac{V_{AC,max}^2}{R_L} \\
 &= \frac{(V_{cc} - v_{ce,sat})^2}{8R_L}
 \end{aligned} \tag{10.51}$$

On the other hand, the DC power drawn from the power supply is given by:

$$\begin{aligned}
 P_{DC} &= V_{cc} I_{DC} \\
 &= \frac{V_{cc} (V_{cc} - v_{ce,sat})}{2R_L}
 \end{aligned} \tag{10.52}$$

Hence, the efficiency of a class A power amplifier, such as that shown in Figure 10.8, is given by:

$$\eta = \frac{1}{4} \frac{V_{cc} - v_{ce,sat}}{V_{cc}} \approx 0.25 \tag{10.53}$$

The last approximation is valid because usually,  $V_{cc} \gg v_{ce,sat}$ . Equation 10.53 indicates that the maximum efficiency of a class A power amplifier, having the configuration shown in Figure 10.8, is 25%. To increase the efficiency of the class A power amplifier an inductor is added in parallel to the load as shown in Figure 10.10.

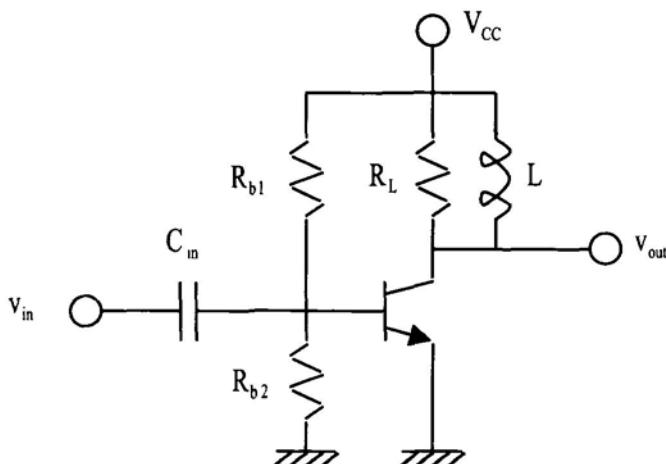


Figure 10.10. Class A power amplifier.

The load impedance of the circuit shown in Figure 10.10 depends on the frequency of the input signal. The load impedance is given by:

$$Z_L = \frac{wLR}{\sqrt{R^2 + w^2L^2}} \quad (10.54)$$

Hence, the circuit of Figure 10.10 has two load lines, a DC load line and an AC load line, as shown in Figure 10.11. Assuming that the inductor has a zero DC resistance, then the DC load impedance is zero, and the DC load line is vertical as shown in Figure 10.11. The intersection of the DC load line with the transfer characteristics of the transistor is the operating point of the transistor. The AC load line passes through the operating point and has a slope equal to the inverse of the load impedance. The maximum AC amplitude at the output of the amplifier, without clipping, is  $V_{cc} - V_{ce,sat}$ . Hence, the maximum AC power delivered to the load  $R_L$  is given by:

$$\begin{aligned} P_{AC,max} &= \frac{1}{2} \frac{V_{AC,max}^2}{R_L} \\ &= \frac{(V_{cc} - v_{ce,sat})^2}{2R_L} \end{aligned} \quad (10.55)$$

On the other hand, the DC power drawn from the power supply is given by:

$$\begin{aligned} P_{DC} &= V_{cc}I_{DC} \\ &= \frac{V_{cc}^2}{R_L} \end{aligned} \quad (10.56)$$

The efficiency of a class A power amplifier, such as that shown in Figure 10.10, is given by:

$$\eta = \frac{1}{2} \frac{(V_{cc} - v_{ce,sat})^2}{V_{cc}^2} \approx 0.5 \quad (10.57)$$

Again, the last approximation is valid because usually  $V_{cc} \gg v_{ce,sat}$ . Even with a shunt inductor as shown in Figure 10.10, the maximum efficiency a class A power amplifier can achieve is 50%. Another disadvantage of the class A power amplifier is that large power dissipation occurs even if there is no AC input signal. In wireless applications, the power amplifier spends long periods of time in the standby mode with no signal to transmit. This lowers the efficiency of the power amplifier even more.

Class B power amplifiers are designed to have zero power dissipation when there is no AC input signal. This helps improve the efficiency of the power

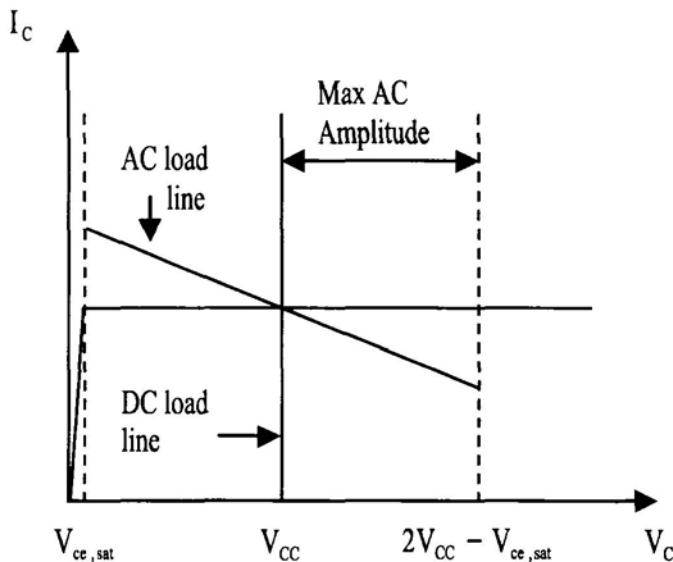


Figure 10.11. Load line of a class A power amplifier.

amplifier during the standby mode or when the input signal is less than the maximum value. The class B power amplifier consists of two complementary transistors (a pnp and an npn transistor), as shown in Figure 10.12. The load resistance is connected to the emitter of the transistors and thus the transistors act as emitter followers. If the input signal is zero, both transistors are off, and the power dissipation of the circuit is zero. When an AC signal is injected into the amplifier, the transistors are on during alternate half cycles. During the positive half cycle  $Q_1$  is on while  $Q_2$  is off. During the negative half cycle  $Q_2$  is on while  $Q_1$  is off. This avoids a direct conduction path between  $V_{cc}$  and ground through the two transistors, and hence improves the efficiency of the power amplifier.

Figure 10.13 shows the transfer characteristics of a class B power amplifier. The transfer characteristics is divided into five regions. In the first region, when  $-v_{be,on} < v_i < v_{be,on}$ , both transistors are not conducting any current. Hence, the current following through the load resistor  $I_L$  is zero, and the output voltage is zero.

The second region is when  $v_{be,on} \leq v_i < V_{cc} + v_{be,on} - v_{ce,sat}$ . In this region,  $Q_1$  is on and acts as an emitter follower in the active region, while  $Q_2$  is off. The output voltage across the load resistor is related to the input voltage by the following equation:

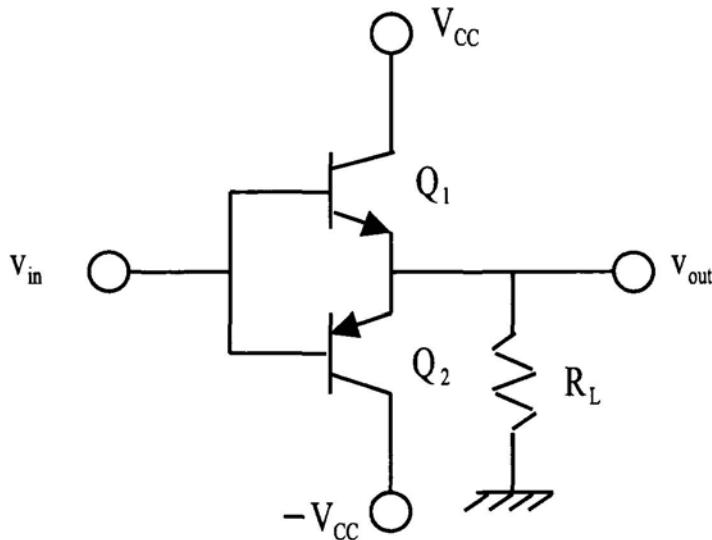


Figure 10.12. Class B power amplifier.

$$v_{out} = v_{in} - v_{be,on} \quad (10.58)$$

The third region is when  $v_i \geq V_{cc} + v_{be,on} - v_{ce,sat}$ . In this region  $Q_2$  remains off, while  $Q_1$  goes into saturation. The output voltage saturates at  $V_{cc} - v_{ce,sat}$ .

The fourth region is when  $-(V_{cc} + v_{be,on} - v_{ce,sat}) < v_i \leq -v_{be,on}$ . In this region,  $Q_2$  is on and acts as an emitter follower in the active region, while  $Q_1$  is off. The output voltage across the load resistor is related to the input voltage by the following equation:

$$v_{out} = v_{in} + v_{be,on} \quad (10.59)$$

The fifth region is when  $v_i \leq -(V_{cc} + v_{be,on} - v_{ce,sat})$ . In this region  $Q_1$  remains off, while  $Q_2$  goes into saturation. The output voltage saturates at  $-(V_{cc} - v_{ce,sat})$ .

Assume that the input signal to the class B power amplifier of Figure 10.12 is given by:

$$v_i = V_{AC} \sin w_c t \quad (10.60)$$

Furthermore, assume that  $V_{AC}$  is much larger than  $v_{be,on}$ . In other words, we are going to neglect the value of  $v_{be,on}$ . During the positive half cycle, the load current is equal to the current flowing through the transistor  $Q_1$  ( $I_{E1}$ ). Therefore,

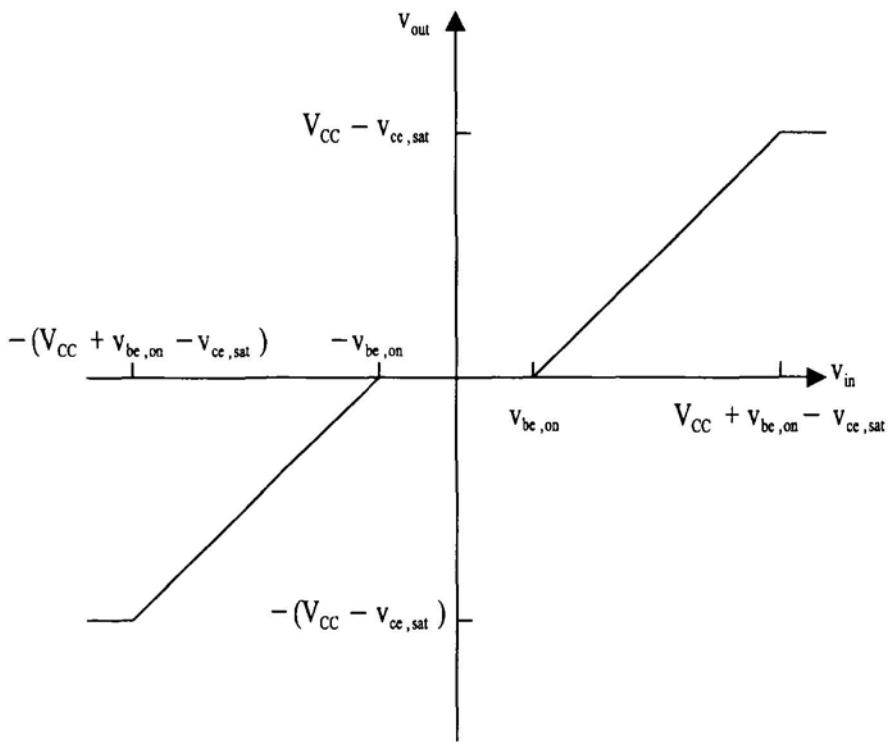


Figure 10.13. Transfer characteristics of a class B power amplifier.

$$I_L = I_{E1} = \frac{V_{AC}}{R_L} \sin w_c t \quad \text{if } v_i > 0 \quad (10.61)$$

The average current flowing through the transistor  $Q_1$ , which equals the average current drawn from  $V_{cc}$  (neglecting the base current), is given by:

$$\bar{I}_{E1} = \frac{1}{\pi} \frac{V_{AC}}{R_L} \quad (10.62)$$

During the negative half cycle, the load current  $I_L$  equals the current flowing through the transistor  $Q_2$  ( $I_{E2}$ ).

$$I_L = I_{E2} = \frac{V_{AC}}{R_L} \sin w_c t \quad \text{if } v_i < 0 \quad (10.63)$$

The average current flowing through the transistor  $Q_2$ , which equals the average current drawn from  $-V_{cc}$  (neglecting the base current), is given by:

$$\bar{I}_{E2} = -\frac{1}{\pi} \frac{V_{AC}}{R_L} \quad (10.64)$$

Therefore, the DC power drawn from the power supply, for the circuit shown in Figure 10.12, is given by:

$$\begin{aligned} P_{DC} &= V_{cc}\bar{I}_{E1} - V_{cc}\bar{I}_{E2} \\ &= \frac{2}{\pi} \frac{V_{AC}}{R_L} \end{aligned} \quad (10.65)$$

The AC power delivered to the load resistance is given by:

$$P_{AC} = \frac{1}{2} \frac{V_{AC}^2}{R_L} \quad (10.66)$$

Hence, the efficiency of the class B power amplifier is given by:

$$\eta = \frac{\pi}{4} \frac{V_{AC}}{V_{CC}} \quad (10.67)$$

Assuming that  $v_{ce,sat}$  is small, the maximum value of  $V_{AC}$  the class B power amplifier can deliver at its output is  $V_{CC}$ . Hence, the maximum efficiency of the class B power amplifier is:

$$\eta_{max} = \frac{\pi}{4} \approx 78\% \quad (10.68)$$

The class B power amplifier has two advantages over the class A power amplifier:

1. It has higher efficiency. The efficiency of the class B power amplifier can reach 78%, while the maximum efficiency of the class A power amplifier is 50%.
2. In standby mode, when the power amplifier has no signal to amplify, the DC power dissipation is zero. This is very useful in wireless applications.

However, class B power amplifiers suffer from crossover distortion. This occurs when the input signal is crossing from the positive half cycle to the negative one, or vice versa. During crossover both devices are off across a finite input voltage range. This can be clearly seen from the transfer characteristics of Figure 10.13. Crossover distortion degrades the linearity of the output waveform.

In a class B power amplifier the transistors are biased at the edge of conduction, so that during the positive half cycle, one transistor is conducting, while during the negative half cycle the other transistor is conducting. Hence, the conduction angle for each transistor is 180°. Taking this principle one step further, it is possible to have the transistor biased below threshold as shown in Figure 10.14. In this case, only during a portion of the positive cycle of the

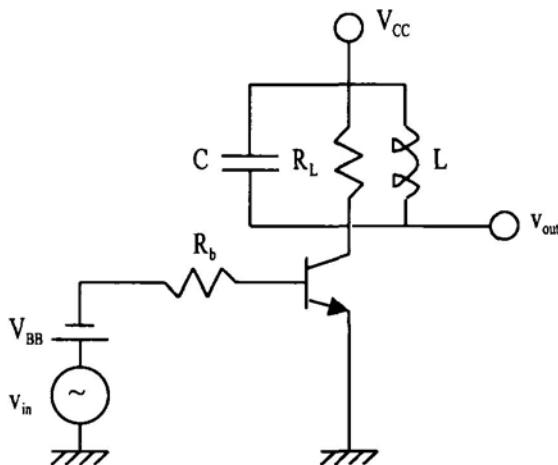


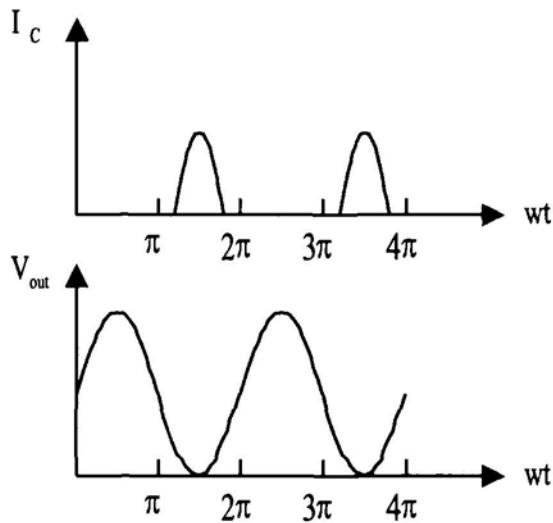
Figure 10.14. Class C Power Amplifier.

input signal, when the input signal is large enough, will the transistor conduct. Hence, the conduction angle is less than  $180^\circ$ .

The pulsed nature of the operation of the transistor makes the relation between the input and output voltages non-linear. This makes the class C power amplifier suitable for constant envelope modulation schemes. Class C power amplifiers can achieve efficiencies in the range of 60% or higher [103]. Figure 10.15 shows the collector current and output voltage waveforms for a class C power amplifier. The reason for the high efficiency is that when the voltage across the transistor is high, the current following through it is small (or zero) as shown in Figure 10.15. On the other hand, when a current flows through the transistor the output voltage is low.

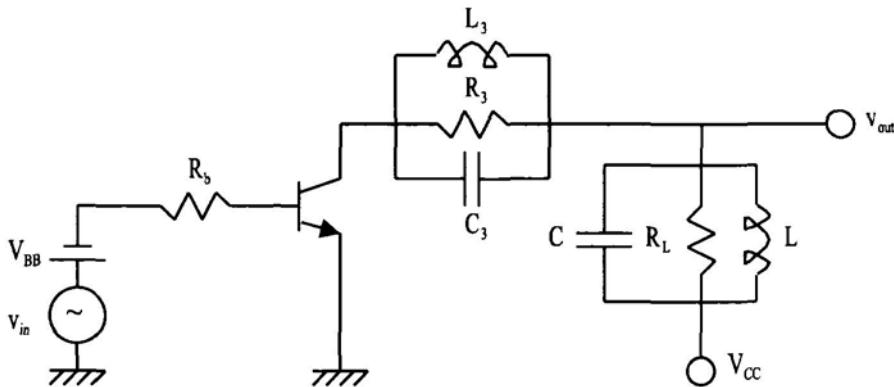
Notice that, in Figure 10.15 when  $I_c$  has a non-zero value,  $v_{out}$  is small but non-zero, except at one point when  $I_c$  is at its peak. To maximize the efficiency of the power amplifier, the voltage across the transistor needs to be more square, such that when the collector current is non-zero, the collector voltage is zero and no power is dissipated in the transistor.

Figure 10.16 shows a class F power amplifier. This is similar to a class C power amplifier with an extra LC tank circuit placed between the transistor and the fundamental LC tank circuit. This LC tank circuit is tuned to the third harmonic of the output signal. The voltage across the transistor is the sum of the fundamental component and the third harmonic. These are the first (and largest) two harmonics of a square wave. Thus, the voltage across the transistor is closer to zero when a current is following through it. Figure 10.17 shows the voltage and current waveforms for a class F power amplifier. This increases



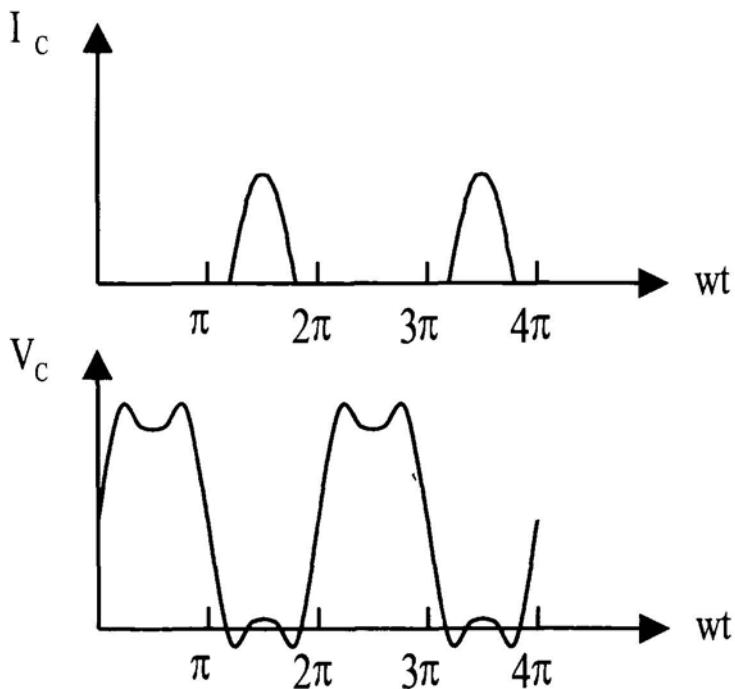
*Figure 10.15.* Waveforms of the collector current and output voltage for a class C power amplifier.

the efficiency of the class F power amplifier over that of the class C power amplifier.



*Figure 10.16.* Class F power Amplifier.

There are practical considerations that need to be considered when designing a power amplifier. While shrinking the process size enables the device to operate at a higher frequency, yet it causes the break down voltage of the device to decrease. This in turn places a limit on the supply voltage. The break



*Figure 10.17. Collector current and voltage waveforms for a class F power amplifier.*

down voltage needs to be three times the supply voltage. Hence, a compromise is required when selecting the transistor size.

The power amplifier is the most power-consuming block in a wireless system. This can cause the power amplifier to over heat, which eventually leads to its failure. To prevent this from happening, the thermal design of the power amplifier needs to be considered carefully. The thermal resistance of a package indicates how poorly the package radiates the heat generated. In general, the larger the thermal resistance, the hotter the package will get. In some cases, to lower the thermal resistance a heat sink is needed. This allows the power amplifier to dissipate more power without significantly increasing its junction temperature.

# Chapter 11

## PHASE LOCKED LOOPS

### 11.1 INTRODUCTION

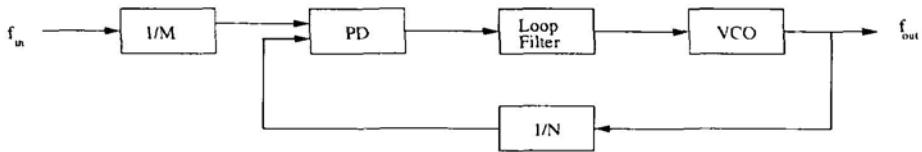
A phase locked loop (PLL) is a device capable of generating an output signal having a frequency that is synchronized to that of an input reference signal. The generated frequency has good noise performance and is of high accuracy. A PLL is used for locking a noisy and imprecise signal to a more precise and less noisy signal, which usually has a lower frequency. Furthermore, a PLL can be implemented on a monolithic chip making it suitable for wireless systems, where size and power dissipation are of high importance. PLLs are extensively used in frequency synthesizers. About 98% of the frequency synthesizer market is based on PLL technology.

The PLL suppress the voltage controlled oscillator (VCO) noise within the bandwidth of its loop, any VCO noise outside the bandwidth of the loop passes virtually unattenuated.

The focus of this chapter is to present different PLL architectures, we also discuss the important PLL parameters. In section 11.2, we present the basic structure of the PLL, and its theory of operation. In section 11.3, we consider different types of phase detectors and their impact on the performance of the phase locked loop. In section 11.4, we present different types of frequency dividers. Section 11.5 discusses the theory of operation of oscillators along with some examples of oscillator circuits.

### 11.2 OPERATION OF THE PHASE LOCKED LOOP

A phase locked loop, as shown in Figure 11.1, typically consists of a phase detector (PD), a loop filter, a voltage controlled oscillator (VCO), and frequency dividers.



*Figure 11.1. Block diagram of a phase locked loop (PLL).*

Now the operation of each component is described. The phase detector has two input signals, and produces an output whose dc level is proportional to the phase difference between these two signals. Alternatively, the operation of the phase detector may be described as:

$$V_{PD}(t) = K_d \phi_e \quad (11.1)$$

Where,

$V_{PD}(t)$  is the output voltage of the phase detector.

$K_d$  is the gain of the phase detector. The unit of  $K_d$  is  $\text{V rad}^{-1}$ .

$\phi_e$  is the difference in phase between the two input signals to the phase detector (measured in radians).

In section 11.3, we discuss some different types of phase detectors and their impact on the performance of the PLL.

The loop filter is a low-pass filter that eliminates the high frequency components leaving only the low frequency ones. The purpose of the loop filter is to smoothen the output of phase detector. The loop filter is also used to control the bandwidth of the closed loop system. The loop filter in most cases is a first or second order low pass RC passive filter.

The voltage-controlled oscillator is a device that produces an output signal having a frequency that is dependent on its input voltage. Typically, the VCO produces a frequency proportional to its input voltage. More specifically:

$$f_{osc}(t) = f_o + \frac{K_o}{2\pi} v_{in}(t) \quad (11.2)$$

Where,

$f_{osc}(t)$  is the output frequency of the VCO.

$f_o$  is the free oscillation frequency (frequency at which the VCO oscillates when no input voltage is applied) measured in Hz.

$K_o$  is the VCO gain (in  $\text{rad s}^{-1} \text{V}^{-1}$ ).

$v_{\text{in}}(t)$  is the input voltage to the VCO.

The voltage range of the VCO is designed so that its output frequency range is at least as large as that required by the PLL.

The frequency dividers allow the input frequency to be different from the output frequency by a factor of  $N/M$ . Note that, if the value of  $N$  or  $M$  is variable, then the PLL-based frequency synthesizer can produce a variable number of output frequencies.

The operation of the PLL is actually quite simple. If the input signal's phase is lagging behind the phase produced by the VCO, this means that the VCO frequency is too high and hence must be lowered. The phase detector handles this by producing an inhibiting signal to the loop filter, which in turn lowers the input voltage to the VCO. As can be shown by equation (11.2), lowering the input voltage to the VCO amounts to lowering the output frequency of the VCO, which is what is desired. On the other hand, if the input signal's phase is leading that of the VCO, the phase detector produces an excitatory signal to the loop filter, which increases the output voltage of the loop filter, and in turn, increases the output frequency of the VCO, as desired.

The PLL tunes the frequency of the voltage-controlled oscillator until the two frequencies at the input to the phase detector are equal:

$$\frac{f_{\text{in}}}{M} = \frac{f_{\text{out}}}{N} \quad (11.3)$$

When this happens, the PLL is said to be locked. Solving equation 11.3 for  $f_{\text{out}}$ , we get:

$$f_{\text{out}} = \frac{N}{M} f_{\text{in}} \quad (11.4)$$

The loop filter is typically a low-pass filter. For a first order low-pass filter, the transfer function is given by:

$$H_{LF}(jw) = \frac{K_f}{1 + jw\tau_f} \quad (11.5)$$

Where,

$K_f$  is the dc gain of the low-pass filter.

$\tau_f$  is the time constant of the low-pass filter.

The input to the voltage-controlled oscillator is a voltage signal  $v_1$ , while the output from the VCO is a signal having a phase  $\phi_o$ .  $\phi_o$  is the integration of  $w_{\text{osc}} = 2\pi f_{\text{osc}}$  over time. The transfer function of the VCO is given by:

$$\frac{\phi_o(jw)}{v_1(jw)} = \frac{K_o}{jw} \quad (11.6)$$

Where,  $K_o$  is the VCO constant in  $\text{rad s}^{-1} \text{V}^{-1}$ . The phase detector produces an output voltage proportional to the phase difference between its two inputs. Hence, the phase detector can be modeled as a subtractor with gain  $K_d$ , where  $K_d$  is the phase detector constant in  $\text{V rad}^{-1}$ .

The system performance of the PLL is also important. Although the behavior of the PLL is highly non-linear, it may be linearized in steady-state conditions [104]. Figure 11.2 shows the block diagram of the PLL in the frequency domain. Solving for the phase transfer function we get:

$$\left( \frac{\phi_i(jw)}{M} - \frac{\phi_o(jw)}{N} \right) K_d H_{LP}(jw) \frac{K_o}{jw} = \phi_o(jw) \quad (11.7)$$

Therefore,

$$H_{PLL}(jw) = \frac{\frac{\phi_o(jw)}{\phi_i(jw)}}{\frac{\frac{K_o K_d}{jw M} H_{LP}(jw)}{1 + \frac{K_o K_d}{jw N} H_{LP}(jw)}} \quad (11.8)$$

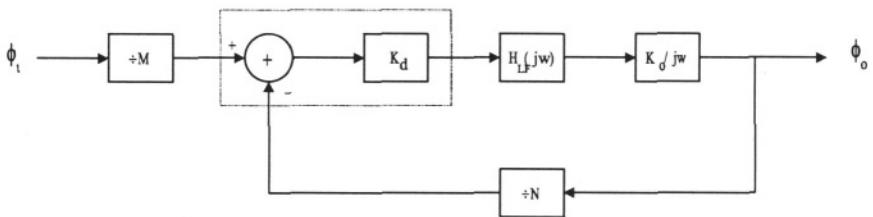


Figure 11.2. Block diagram of a phase locked loop (PLL) in the frequency-domain.

Assuming the low-pass filter is a first-order low-pass filter having a transfer function as given by equation 11.5, hence the transfer function of the PLL becomes:

$$H_{PLL}(jw) = \frac{N}{M} \frac{1}{\frac{N \tau_f}{K_o K_d K_f} (jw)^2 + \frac{N}{K_o K_d K_f} jw + 1} \quad (11.9)$$

Clearly this is a second order system. This is advantageous since second order systems are well-known and have been analyzed extensively in the past. The transfer function  $H(jw)$  can be expressed as:

$$H_{PLL}(jw) = \frac{K_{dc}}{\left(\frac{jw}{w_n}\right)^2 + \frac{2\xi}{w_n}jw + 1} \quad (11.10)$$

Where,

$K_{dc}$  is the dc gain of the second order system.

$w_n$  is the natural frequency of the PLL. This is related to the PLL parameters by the following equation:

$$w_n^2 = \frac{K_o K_d K_f}{N \tau_f} \quad (11.11)$$

$\xi$  is the damping factor of the PLL [105]. This is related to the PLL parameters by the following equation:

$$\xi = \frac{1}{2} \sqrt{\frac{N}{K_o K_d K_f \tau_f}} \quad (11.12)$$

Both the natural frequency and the damping factor are critical in determining the overall frequency response and time response of a second order system. Depending on the damping factor ( $\xi$ ) the system can be in one of three regions.

If  $\xi > 1$ , the system is said to over-damped. In this case, the step response of the second order system is given by:

$$u_{PLL}(t) = K_{dc} \left\{ 1 - e^{-w_n \xi t} \left[ \cosh \left( w_n \sqrt{\xi^2 - 1} t \right) + \frac{\xi}{\sqrt{\xi^2 - 1}} \sinh \left( w_n \sqrt{\xi^2 - 1} t \right) \right] \right\} \quad (11.13)$$

If  $\xi = 1$ , the system is said to be critically damped. In this case, the step response of the second order system is given by:

$$u_{PLL}(t) = K_{dc} \left\{ 1 - e^{-w_n t} [1 + w_n t] \right\} \quad (11.14)$$

If  $0 < \xi < 1$ , the system is said to be under-damped. In this case, the step response of the second order system is given by:

$$u_{PLL}(t) = K_{dc} \left\{ 1 - e^{-w_n \xi t} \left[ \cos(w_n \sqrt{1 - \xi^2} t) + \frac{\xi}{\sqrt{1 - \xi^2}} \sin(w_n \sqrt{1 - \xi^2} t) \right] \right\} \quad (11.15)$$

Figure 11.3 shows the step response of the phase locked loop, as a function of normalized time ( $w_n t$ ), for different values of  $\xi$ . Notice that, in the under-damped case when,  $\xi < 1$ , there are over-shoots in the step response. On the other hand, if  $\xi > 1$ , the rise time of the step response becomes relatively long.

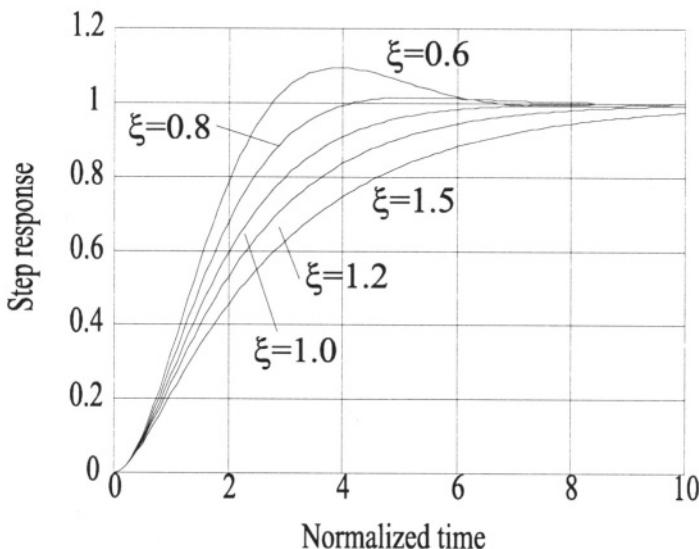


Figure 11.3. Step response of a phase locked loop.

Another important system level issue in PLLs is its stability. Since the phase-locked loop contains a feedback path it effectively becomes a control system. In control systems it is customary to represent the transfer function by  $H(s)$  instead of  $H(j\omega)$ . For stability, the poles must be on the left hand side of the complex plane. This means that the roots of the transfer function,  $H(s)$ , must both have negative real parts. The poles of  $H(s)$  may be given as:

$$p_1, p_2 = \xi \omega_n \pm w_n \sqrt{\xi^2 - 1} \quad (11.16)$$

This means that both  $\text{Re}(p_1)$  and  $\text{Re}(p_2)$  must be negative in order for the PLL to be stable.

## 11.3 PHASE DETECTORS

The phase detector, also known as a phase comparator, is a device that compares the phase of two signals. Ideally, the output of the phase detector should be proportional to the phase difference between its input signals:

$$V_{PD} = K_d(\phi_1 - \phi_2) \quad (11.17)$$

Where,

$V_{PD}$  is the output voltage of the phase detector.

$K_d$  is the phase detector constant in volts/radian.

$\phi_1$  and  $\phi_2$  are the phase of the two input signals.

In practice, the phase detector is a non-linear device and equation 11.17 is only an approximation to reality. There are many types of phase detectors that are used in PLLs, some of the typical architectures used are explained in this section.

### 11.3.1 Analog Phase Detectors

The most popular analog phase detector is the multiplier phase detector which produces an output signal that is proportional to the product of the two input signals. Assume that the two input signals are given by:

$$v_1(t) = V_1 \cos[w_1 t + \phi_1(t)]$$

$$v_2(t) = V_2 \cos[w_2 t + \phi_2(t)]$$

Hence, the output of the phase detector is given by:

$$\begin{aligned} V_{PD}(t) &= K_d v_1(t) \cdot v_2(t) \\ &= \frac{K_d V_1 V_2}{2} \{ \cos[(w_1 - w_2)t + \phi_1(t) - \phi_2(t)] \\ &\quad + \cos[(w_1 + w_2)t + \phi_1(t) + \phi_2(t)] \} \end{aligned} \quad (11.18)$$

Where,  $K_d$  is the phase detector constant. The output of the phase detector is filtered such that the component representing the sum of the two phases in equation 11.18, which has a frequency that is equal to the sum of frequencies of the two input signals, is eliminated. Hence, the output of the phase detector becomes:

$$V_{PD}(t) = \frac{K_d V_1 V_2}{2} \cos[(w_1 - w_2)t + \phi_1(t) - \phi_2(t)] \quad (11.19)$$

If  $(\phi_1(t) - \phi_2(t)) \approx \pi/2$ , equation 11.19 can be linearized to:

$$V_{PD}(t) \approx \frac{K_d V_1 V_2}{2} \left\{ \frac{\pi}{2} + (w_1 - w_2)t + \phi_1(t) - \phi_2(t) \right\} \quad (11.20)$$

For the component representing the difference between the two phases of the input signals to pass unfiltered through the loop filter, the following condition needs to be satisfied:

$$\frac{1}{2\pi}(w_1 - w_2) < \text{BW} \quad (11.21)$$

Where, BW is the bandwidth of the loop filter. If condition 11.21 is satisfied, the PLL is said to be in the lock range and becomes locked ( $w_1 = w_2$ ) in a short period of time.

Once the PLL becomes locked, the output of the loop filter following the phase detector becomes:

$$V_{LF} = \frac{K_d V_1 V_2}{2} \cos[\phi_1 - \phi_2] \quad (11.22)$$

Equation 11.22 is plotted in Figure 11.4. As long as the voltage needed to set the VCO to make  $w_1 = w_2$  doesn't exceed  $\frac{K_d V_1 V_2}{2}$ , which is the maximum voltage the loop filter can produce, the PLL is said to be in the hold range.

The multiplier phase detector can have excellent (very low) noise levels, which is hard to achieve using digital phase detectors. The reason for this being that we average the phase shift over the entire period and not just rely on the zero crossing points as in the digital phase detectors.

### 11.3.2 Digital Phase Detectors

Digital phase detectors are designed to operate on bi-level digital signals. A simple digital phase detector is the exclusive OR (XOR) gate. To understand how the XOR gate operates as a phase detector consider the waveforms shown in Figure 11.5. Assume that the input waveforms to the XOR gate (I1 and I2) have the same frequency and a 50% duty cycle (the duration of the logic “1” state is equal to the duration of the logic “0” state). The phase shift between I1 and I2 is zero in Figure 11.5.a, 1/8 of a cycle ( $\pi/4$ ) in Figure 11.5.b, 1/4 of a cycle ( $\pi/2$ ) in Figure 11.5.c, and half a cycle ( $\pi$ ) in Figure 11.5.d.

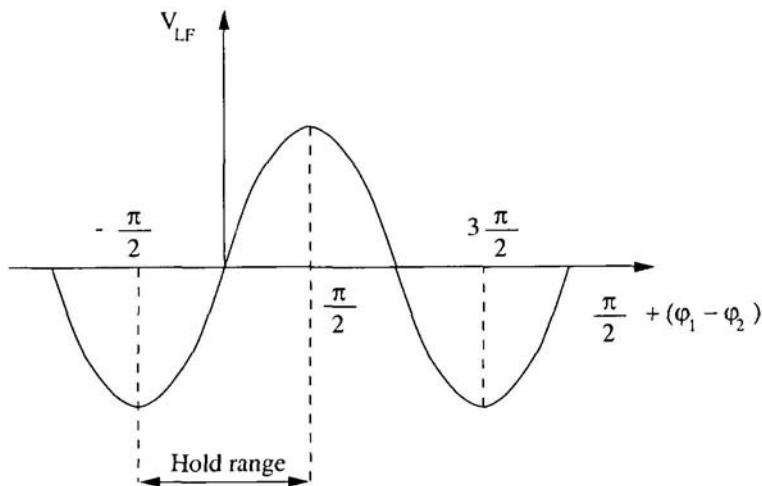
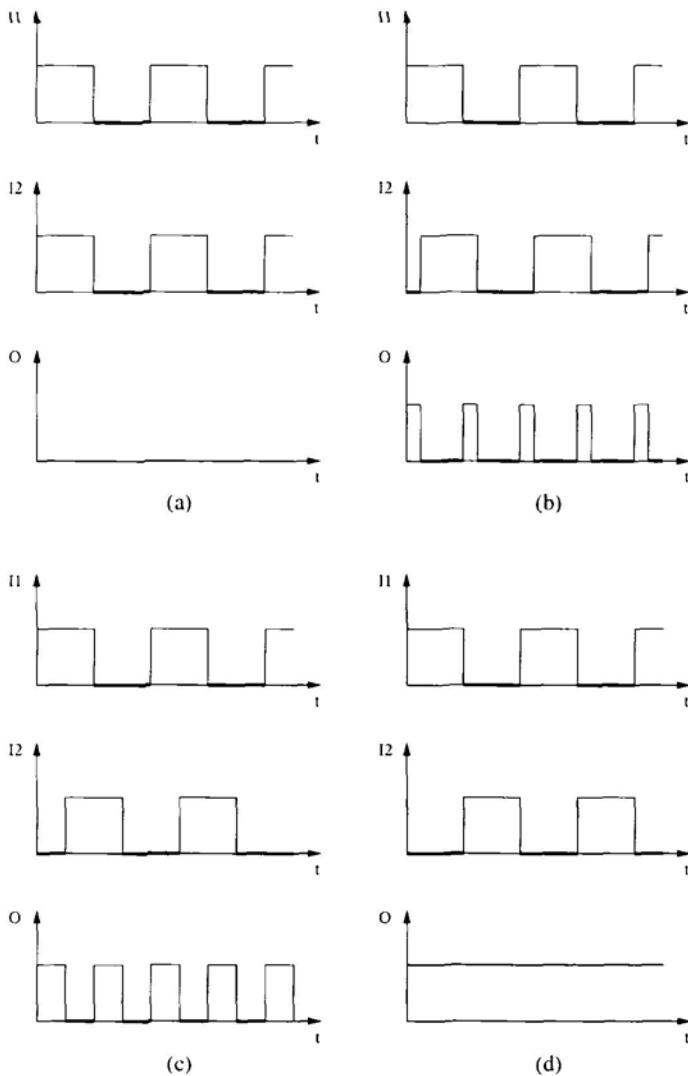


Figure 11.4. Loop filter output versus phase error at the input of a multiplier phase detector.

When the phase shift between the two input waveforms to the XOR phase detector is zero, the output of the phase detector is always logic state zero. As the phase shift between the two inputs starts to increase, the output of the phase detector starts alternating between logic state zero and logic state one. The duty cycle of output increases as the phase shift increases. Hence, the average value of the phase detector output, which is the output of the loop filter increases as well. This is shown in the transfer characteristics of Figure 11.6. The increase in the loop filter output with the increase in phase shift between the two input signals continues until we reach a phase shift of half-a-cycle. At this phase shift, which is shown in Figure 11.5.d, the output of the phase detector is always in logical state one. Any increase in phase shift beyond this point causes the duty cycle of the phase detector output to decrease and hence the average voltage at the output of the phase detector decreases as well, this is shown in Figure 11.6.

Figure 11.7 shows the block diagram of a digital phase detector using edge-triggered flip flops. This phase detector can be represented by a three-state machine as shown in Figure 11.8. The phase detector transitions from one state to the next upon the arrival of a rising edge on one of the inputs. Assume that phase detector is in state two, where both  $Q_1$  and  $Q_2$  are low. A rising edge arrives on the oscillator input, the phase detector transitions to state three, where  $Q_1$  is high and  $Q_2$  is low. A rising edge on the reference input transitions the phase detector back to state two, while a rising edge on the oscillator input keeps the phase detector in state three. Notice that, when the phase detector transitions from state three to state two, it temporary passes through a state where both  $Q_1$  and  $Q_2$  are high. Hence, the output of the AND



*Figure 11.5.* Input and output waveforms to an XOR digital phase detector.

gate is high, this resets both flip flops making the phase detector go to state two.

Similarly, when a rising edge arrives on the reference input and the phase detector is in state two, it transitions to state one, where  $Q_1$  is low and  $Q_2$  is high. A rising edge on the oscillator input transitions the phase detector back to state two, while a rising edge on the reference input keeps the phase detector in state one. Notice that, when the phase detector transitions from state one to state two, it temporary passes through a state where both  $Q_1$  and  $Q_2$  are high.

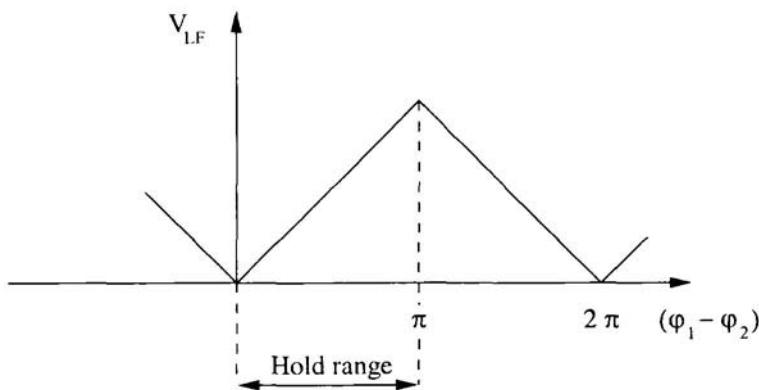


Figure 11.6. Output voltage of the loop filter versus the phase shift between the two inputs to an XOR digital phase detector.

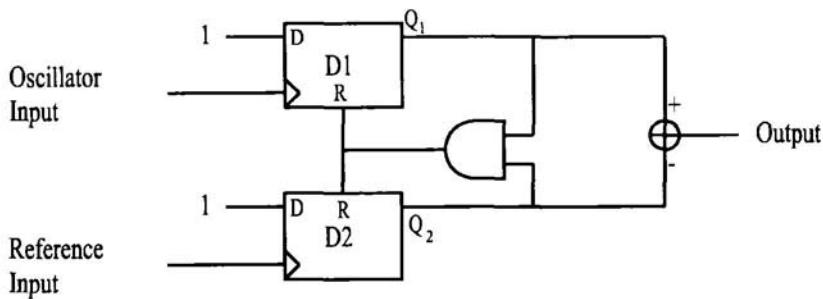


Figure 11.7. Block diagram of a digital phase detector using edge-triggered flip flops.

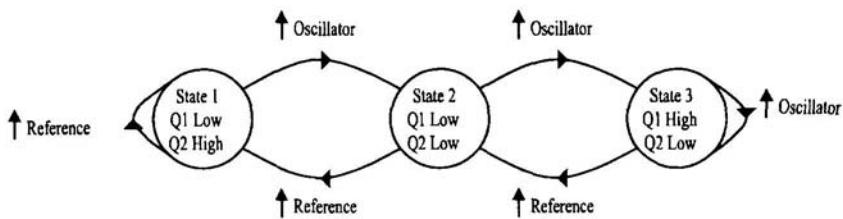
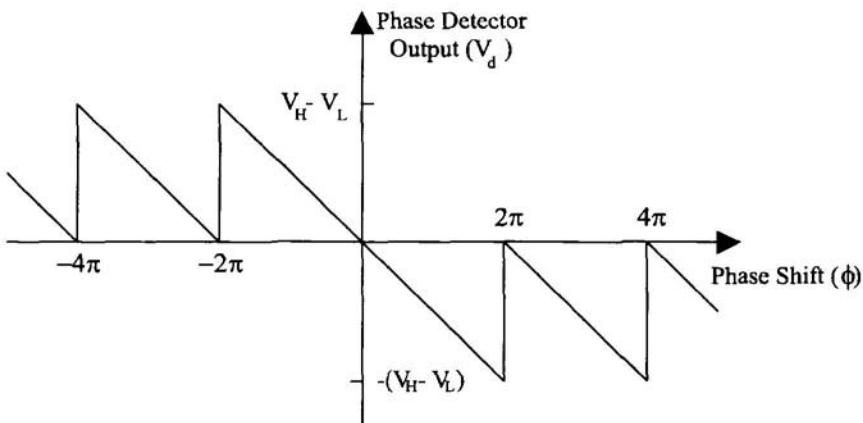


Figure 11.8. State machine representation of the digital phase detector of Figure 11.7.

Hence, the output of the AND gate is high, this resets both flip flops making the phase detector go to state two. The output of the phase detector is zero when it is in state two, positive when it is in state three, and negative when it is in state one.

If the oscillator frequency is higher than the reference frequency, there will be more positive edges on the oscillator input than on the reference input. Therefore, the phase detector transitions between states two and three, and will accordingly have a positive average value on its output. This should steer the VCO towards making the oscillator and reference frequencies equal. Similarly, when the reference frequency is higher than the oscillator frequency, there will be more positive edges on the reference input than on the oscillator input. Therefore, the phase detector transitions between states one and two, and will accordingly have a negative average value on its output. This should steer the VCO towards making the oscillator and reference frequencies equal. Figure 11.9 shows the transfer function of the digital phase detector. Notice that, the this phase detector has a linear region which extends from  $-2\pi$  to  $2\pi$ .



*Figure 11.9.* Transfer function of the digital phase detector of Figure 11.7.

Figure 11.10 shows the block diagram of another digital phase detector using edge-triggered flip flops. The flip flops of the input stage D1 and D2 operate as frequency dividers, they divide the frequency of the reference signal and the oscillator signal by a factor of 2. When the reference and oscillator frequencies are equal, the flip flops of the output stage D3 and D4 are disabled.  $\bar{Q}_3$  and  $Q_4$  are both high (at logical “1”), hence the output of the phase detector is equal to the output of the XOR gate. When the frequencies of the two input signals start shifting with respect to each other,  $\bar{Q}_3$  or  $Q_4$  becomes low (logical “0”), thus overriding the output of the XOR, making the output

of the phase detector saturate at the level that steers the reference and oscillator frequencies towards lock.

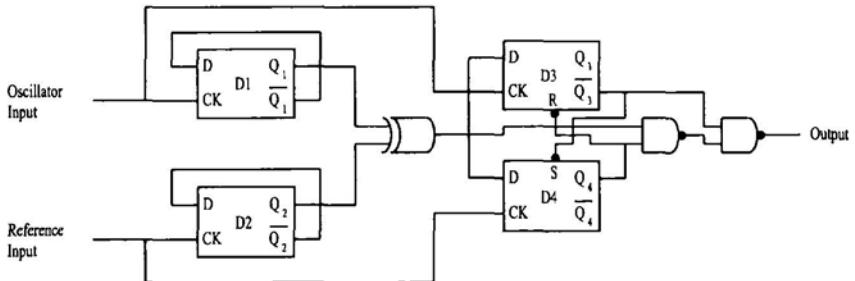


Figure 11.10. Block diagram of an edge-triggered phase detector.

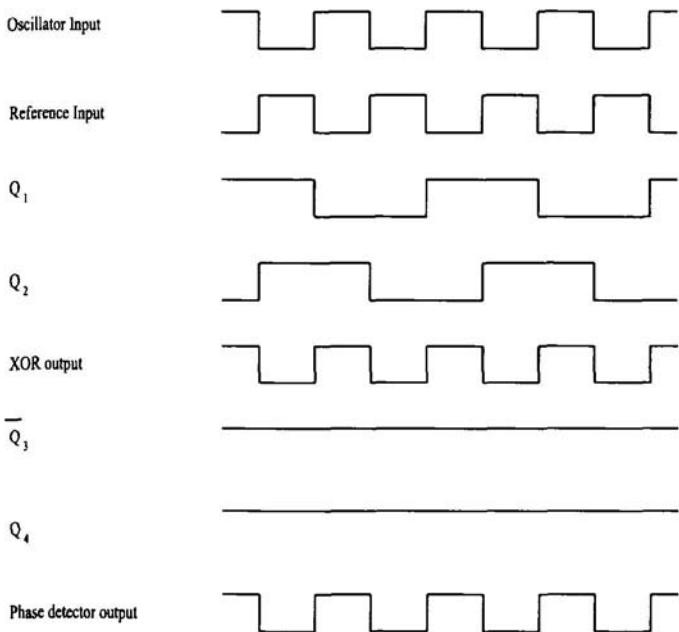
Figure 11.11 shows the timing diagram for the edge-triggered phase detector when the oscillator and reference frequencies are equal but out of phase by  $180^\circ$ . In this case, the outputs of D1 and D2 are out of phase by  $90^\circ$ , and the XOR gate has a rectangular wave output with a 50% duty cycle. Assume that, just before the arrival of the positive edge on the reference input, the output of the XOR gate is high (logical “1”). When the positive edge of the reference input arrives,  $Q_2$  changes its logical state and D4 stores logical “1” making  $Q_4 = 1$ .

Similarly for the oscillator input, just before the arrival of the positive edge on the oscillator input, the XOR gate output is low (logical “0”). When the positive edge of the oscillator input arrives,  $Q_1$  changes its logical state and D3 stores logical “0” making  $\bar{Q}_3 = 1$ .

If the phase shift between the two input signals were to decrease below  $180^\circ$ , the phase shift of the outputs of D1 and D2 decreases below  $90^\circ$ , and the duty cycle of the signal at the output of the XOR gate and the phase detector decreases below 50%. In this case,  $Q_4$  remains high (logical “1”), and  $Q_3$  remains low (logical “0”).

On the other hand, if the phase shift between the two input signals were to increase above  $180^\circ$ , the phase shift of the outputs of D1 and D2 increases above  $90^\circ$ , and the duty cycle of the signal at the output of the XOR gate and the phase detector increases above 50%. In this case also,  $Q_4$  remains high (logical “1”), and  $Q_3$  remains low (logical “0”). Notice that, the region of linear operation of the phase detector extends from 0 to  $2\pi$ , this is double the region of linear operation of the XOR phase detector.

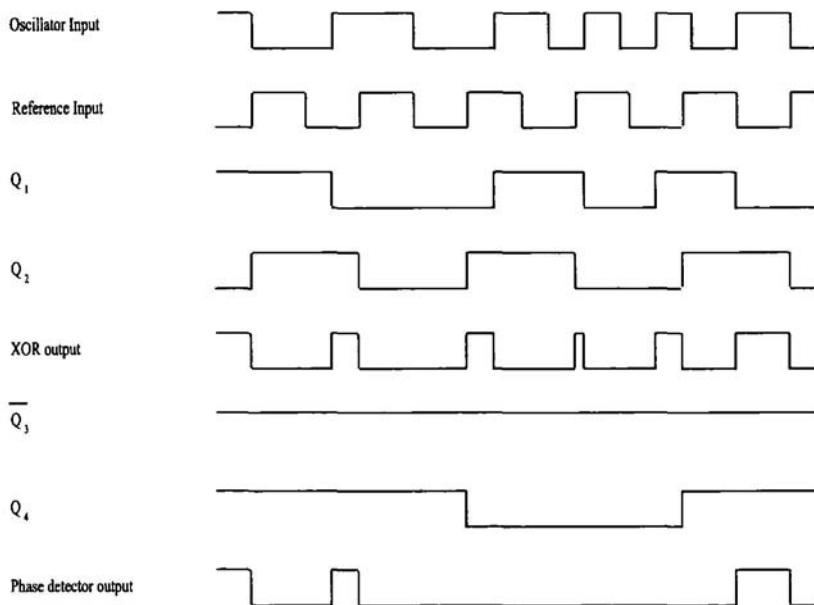
Assume now that reference input has a higher frequency than the oscillator input as shown in Figure 11.12. In this case, it is possible to get two consecutive



*Figure 11.11.* Timing diagram for the edge-triggered phase detector of Figure 11.10, having a  $180^\circ$  phase shift between its inputs.

positive edges on the reference input with no positive edge occurring on the oscillator input. When the second positive edge occurs on the reference input, the output of the XOR gate is logic zero, hence  $Q_4$  goes to logic zero and forces D3 to be in reset ( $Q_3 = 0$ ) regardless of its clock and D input. Since,  $Q_4 = 0$  and  $\overline{Q}_3 = 1$ , therefore the output of the phase detector is low (logic "0"). The phase detector output remains in this state, forcing the oscillator frequency to go to a higher value, until two consecutive positive edges occur on the oscillator input with no positive edge occurring in between on the reference input. In this case, the phase detector returns to the state where  $Q_4 = 1$  and  $\overline{Q}_3 = 1$  and the output of the phase detector is equal to the output of the XOR gate.

Similarly as shown in Figure 11.13, that when oscillator input has a higher frequency than the reference input,  $\overline{Q}_3$  goes to logic 0, forcing D4 to be set ( $Q_4 = 1$ ). Hence, the output of the phase detector goes to logic 1. The phase detector output remains in this state, forcing the oscillator frequency to go to a lower value, until two consecutive positive edges occur on the reference input with no positive edge occurring in between on the oscillator input. In this case,



*Figure 11.12.* Timing diagram for the edge triggered phase detector of Figure 11.10, having higher frequency on the reference input.

the phase detector returns to the state where  $Q_4 = 1$  and  $\overline{Q}_3 = 1$  and the output of the phase detector is equal to the output of the XOR gate.

## 11.4 FREQUENCY DIVIDERS

A frequency divider is used in the phased locked loop to make the frequency of the output signal an integer or fractional multiple of the reference frequency. The dividend can be a fixed division ratio. In this case, the divider can be optimized for high speed and low power operation. A divider having a fixed division ratio is known as a prescaler.

Alternatively, the divider can be a programmable divider, where the division ratio is arbitrary selected by the user. Programmable dividers are used in frequency synthesizers, where the frequency of the output signal of the PLL can be changed based on the divider's division ratio.

A very simple frequency divider is the toggle flip flop (T-flip flop). This is a D-flip flop that has its inverted output ( $\overline{Q}$ ) connected to its input (D), as shown in Figure 11.14.a. This flip flop is capable of dividing the input signal, which

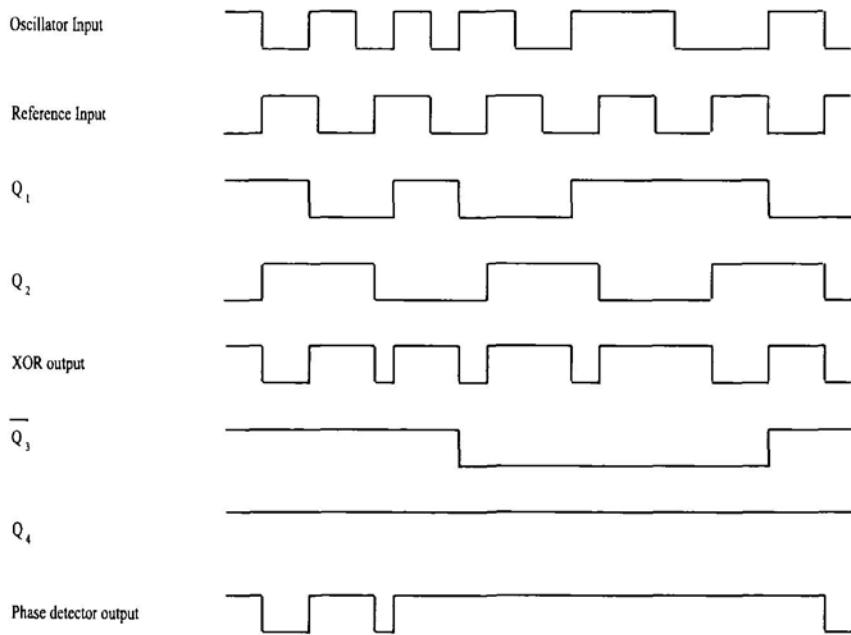


Figure 11.13. Timing diagram for the edge triggered phase detector of Figure 11.10, having higher frequency on the oscillator input.

is connected to the clock input of the flip flop, by a factor of 2. Figure 11.14.b shows the input and output waveforms to this divider.

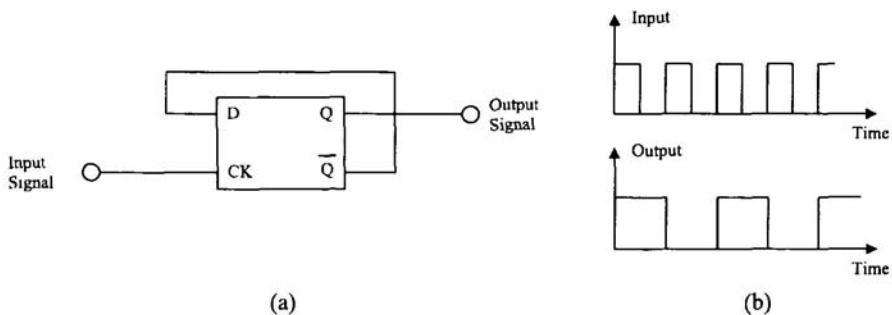


Figure 11.14. Toggle flip flop: (a) Block diagram. (b) Input/output waveforms.

This idea can be extended to a divide-by-four circuit by cascading the output of the first T-flip flop to the input of a second T-flip flop. In general, it is possible to have a divide-by- $2^n$  frequency divider, by having an n-stage ripple counter. An n-stage ripple counter consists of  $n$  T-flip flops connected in cascade as shown in Figure 11.15.

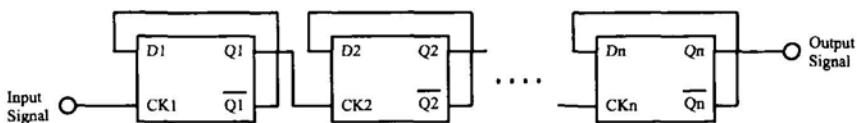


Figure 11.15.  $n$ -stage ripple counter.

Ripple counters of the type shown in Figure 11.15 are limited to division factors that are a power of two. To be able to use any integer as a division factor, synchronous counters such as that shown in Figure 11.16 are used. The counter starts up in the zero state, where the outputs of all the flip flops are zero. On every rising edge of the input signal, the count of the counter is incremented by one. The digital output of the counter, which is the output of its flip flops is compared with the division factor  $N$ , should they be equal, the comparator generates a high logic signal that resets the flip flops to the zero state, and this cycle repeats. Figure 11.17 shows the timing diagram for a divide-by-six counter.

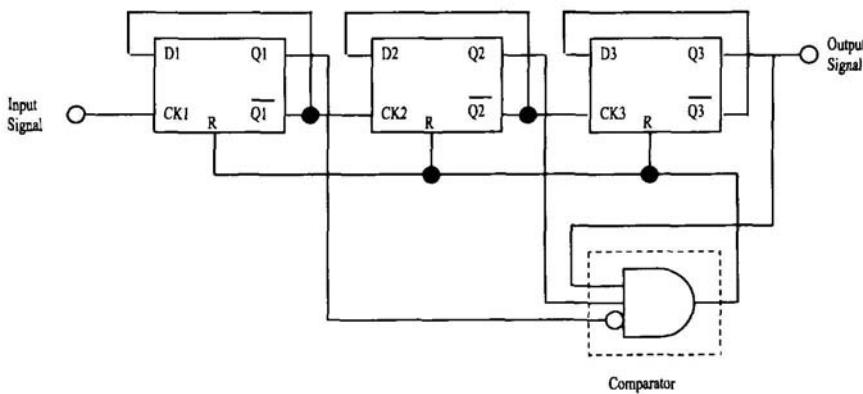


Figure 11.16. A synchronous divide-by-six counter.

The output of the comparator, which can be taken as the output of the frequency divider, has a frequency that is  $1/N$  the input frequency. However, the comparator output has a very small duty cycle, when the comparator output goes high, flip flops are reset forcing the comparator output to be low again. Thus, the comparator output is only momentary in the high state. Notice that, the output of the last flip flop also has a frequency  $1/N$  the input frequency, and it has a 33.33% duty (for the divide-by-six counter of Figure 11.16), hence this signal can be taken as the output of the frequency divider.

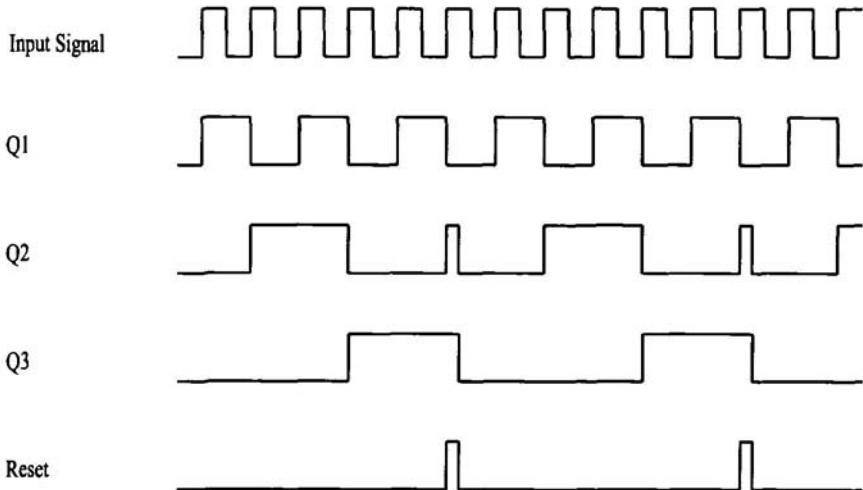


Figure 11.17. Timing diagram for a divide-by-six counter.

The number of flip flops  $n$  needed to implement a synchronous counter depends on the maximum division factor of the frequency divider. If  $N_{\max}$  is the maximum division factor, then  $n$  is given by:

$$n = \lceil \log_2 N_{\max} \rceil \quad (11.23)$$

Where,  $\lceil x \rceil$  is the smallest integer larger than or equal to  $x$ .

A programmable comparator can be implemented using  $n$  XNOR gates and one AND gate as shown in Figure 11.18. Assume that  $C = C_{n-1}C_{n-2}\dots C_1C_0$  is the digital output of the counter, and  $R = R_{n-1}R_{n-2}\dots R_1R_0$  is the desired division factor. Whenever,  $R = C$ , the output of all the XNOR gates is “1”, hence the output of the comparator goes high. This resets the counter to the zero state.

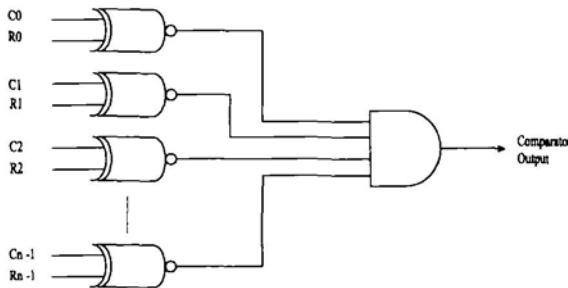


Figure 11.18. A programmable comparator.

## 11.5 OSCILLATOR DESIGN

The oscillator is a periodic waveform generator. It converts DC voltage into a periodic signal. Hence, oscillators are DC-to-RF converters. If the generated waveform is sinusoidal signal the oscillator is said to be a linear oscillator, such oscillators are also known as harmonic oscillators. If the generated waveform has any other shape (e.g. a square wave) the oscillator is said be a nonlinear oscillator [106], such oscillators are also known as relaxation oscillators.

Most oscillators use positive feedback to generate oscillations. A block diagram of a positive feedback oscillator is shown in Figure 11.19. The oscillator consists of two main components: An active element, which acts as an amplifier ( $A$ ), and a passive frequency-dependent feedback network ( $\beta$ ). At the frequency of oscillation the loop-gain  $\beta A$  is equal to unity. At this frequency, the magnitude of the loop-gain is unity and phase shift of the loop-gain is zero, this condition is known as the Barkhausen criterion. The feedback amplifier of Figure 11.19 has a gain of [107]

$$\begin{aligned} A_F &= \frac{X_o}{X_i} \\ &= \frac{A}{1 - \beta A} \end{aligned} \quad (11.24)$$

When  $\beta A = 1$ ,  $A_F$  becomes infinity, which means it is possible to have a sinusoidal output with no input. The frequency of oscillation is determined by the  $\beta$  network. The  $\beta$  network has some type of resonator which makes  $\beta A$  equal to unity only at a single frequency, this is the frequency of oscillation.

When the oscillator is turned on, there is no input signal except for noise. The noise is amplified by the amplifier and then filtered by the  $\beta$  network to select the desired frequency of oscillation. The output of the  $\beta$  network is then

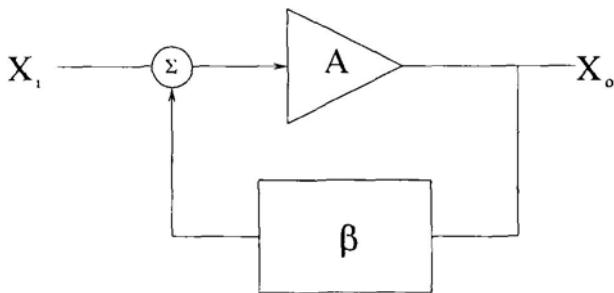


Figure 11.19. Block diagram of a positive feedback oscillator

feedback into the amplifier and this process repeats. Initially, the loop gain is greater than unity, which means that the signal is amplified every time it passes through the loop. As the signal gets larger, the gain of the amplifier gets lower because of the non-linearity. Eventually the loop gain becomes unity and the amplitude of the output oscillations is sustained.

Despite the reliance on the nonlinear characteristics of the active device to stabilize the output, yet the output sinusoidal signal can be of high purity because of the filtering action of the  $\beta$  network. The  $\beta$  network filters out the harmonics of the desired oscillation frequency.

### 11.5.1 Phase-Shift Oscillator

Figure 11.20 shows the circuit diagram of a phase-shift oscillator [107] using a FET as an active element. The amplifier of the phase-shift oscillator can be a common-source FET or a common-emitter BJT. Both amplifier configurations are inverting amplifiers. The phase shift through the amplifier is  $180^\circ$ . The  $\beta$  network of the phase-shift oscillator is a cascade of RC sections. Each RC stage is capable of providing a phase shift between  $0^\circ$  to  $90^\circ$ . To have a unity loop-gain, the  $\beta$  network must be capable of providing a  $180^\circ$  phase-shift, for this to happen a minimum of 3 RC sections is required. With two RC sections a phase shift of  $180^\circ$  can only be obtained as the frequency approaches zero.

The transfer function of a  $\beta$  network consisting of 3 RC sections (as shown in Figure 11.20) is given by:

$$\beta = \frac{(wRC)^3}{wRC(5 - (wRC)^2 + j(6(wRC)^2 - 1)} \quad (11.25)$$

To make the phase shift of the  $\beta$  network equal to  $180^\circ$ , the imaginary part must equal zero. Hence, the frequency of oscillation is:

$$f_{osc} = \frac{1}{2\pi RC\sqrt{6}} \quad (11.26)$$

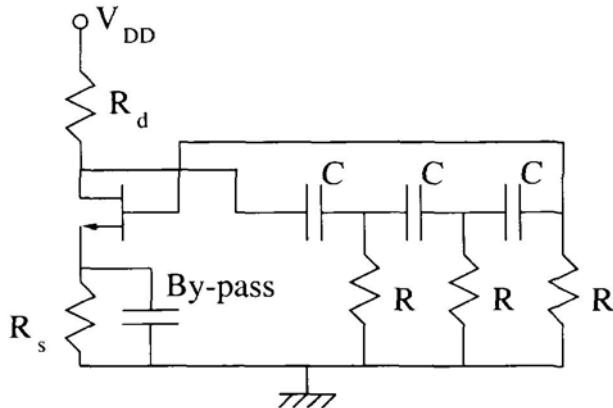


Figure 11.20. Circuit diagram of a phase shift oscillator

The gain, or rather the attenuation, of the  $\beta$  network becomes  $-1/29$ . Therefore, to be able to sustain oscillations, the gain of the FET amplifier ( $|A|$ ) must be greater than 29.

### 11.5.2 Colpitts and Hartley Oscillators

The general form of the Colpitts/Hartley oscillator is shown in Figure 11.21. The active device is a three-terminal FET or BJT (in the good old days vacuum tubes were also used). The  $\beta$  network is a  $\pi$  network consisting of inductors and capacitors.

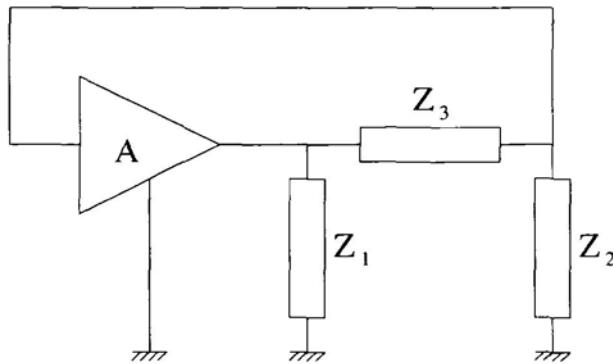


Figure 11.21. Generalized circuit diagram of the Colpitts/Hartley Oscillator.

The active device can be modeled as shown in Figure 11.22. For a FET, the input resistance  $R_i = \infty$ . The output resistance is the drain resistance,  $R_o = r_d$ . For a BJT, the input resistance is the base resistance,  $R_i = r_b$ .

The output resistance  $R_o = \infty$ . The transconductance  $g_m$  of the BJT is  $\beta/r_b$ , where  $\beta$  is the ratio between the collector current to the base current.

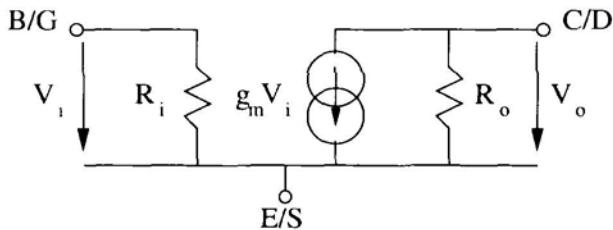


Figure 11.22. Small signal model for active devices. B, E, C are the base, emitter, and collector terminals of a BJT. G, S, D are the gate, source and drain terminals of a FET.

Replacing the active device of Figure 11.21 by its small signal model we get the equivalent circuit of Figure 11.23. The oscillation condition means that  $V_i$  and  $V_o$  exist (have nonzero values) when there is no external excitation.

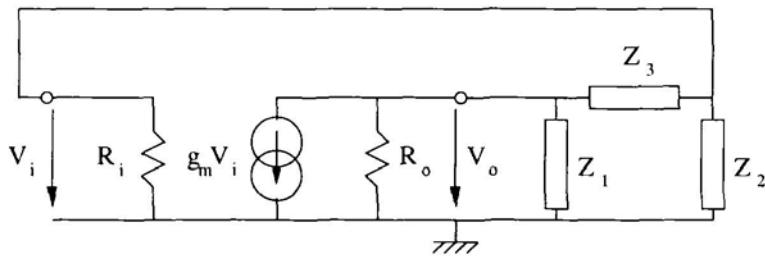


Figure 11.23. Generalized equivalent circuit for the Colpitts/Hartley Oscillator.

$V_i$  and  $V_o$  of Figure 11.23 are inter-related through the active device and the  $\beta$  network. By solving these two equations together we can determine the two conditions of oscillation. For the active device:

$$V_o = -g_m V_i \quad [R_o \parallel Z_1 \parallel (Z_3 + Z_2 \parallel R_i)] \quad (11.27)$$

For the  $\beta$  network:

$$V_i = V_o \frac{Z_2 \parallel R_i}{Z_3 + Z_2 \parallel R_i} \quad (11.28)$$

Solving equations (11.27) and (11.28), we get the condition of oscillation:

$$\frac{Z_3 + Z_2 \parallel R_i}{Z_2 \parallel R_i} = -g_m \quad [R_o \parallel Z_1 \parallel (Z_3 + Z_2 \parallel R_i)]$$

This equation can be simplified into:

$$g_m [Z_2 \parallel R_i] = -\left(\frac{Z_3 + Z_2 \parallel R_i}{Z_1 \parallel R_o} + 1\right) \quad (11.29)$$

Alternatively, equation 11.29 can be found by the writing the nodal equations for Figure 11.23. For the input node:

$$\left[\frac{1}{R_i} + \frac{1}{Z_2} + \frac{1}{Z_3}\right] V_i - \frac{V_o}{Z_3} = 0 \quad (11.30)$$

For the output node:

$$\left[g_m - \frac{1}{Z_3}\right] V_i + \left[\frac{1}{R_o} + \frac{1}{Z_1} + \frac{1}{Z_3}\right] V_o \quad (11.31)$$

Equations 11.30 and 11.31 can be written in matrix form as:

$$\begin{bmatrix} \frac{1}{R_i} + \frac{1}{Z_2} + \frac{1}{Z_3} & -\frac{1}{Z_3} \\ g_m - \frac{1}{Z_3} & \frac{1}{R_o} + \frac{1}{Z_1} + \frac{1}{Z_3} \end{bmatrix} \begin{bmatrix} V_1 \\ V_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad (11.32)$$

For a non-trivial solution to exist for equation 11.32, the determinate of the matrix must equal zero. This leads to the condition of oscillation given by equation 11.29.

$Z_1, Z_2$  and  $Z_3$  are reactive elements.  $Z_n = jX_n$ , for  $n = 1, 2, 3$ . Equating the real and imaginary parts of equation 11.29, we get the following conditions:

$$\frac{X_1 X_2 X_3}{R_i R_o} = (X_1 + X_2 + X_3) \quad (11.33)$$

$$\frac{X_1 X_3}{R_o} + X_1 X_2 \left(\frac{1}{R_i} + \frac{1}{R_o} + g_m\right) + \frac{X_2 X_3}{R_i} = 0 \quad (11.34)$$

For a BJT,  $R_o = \infty$  and  $g_m R_i = \beta$ . Hence, the conditions of oscillation (equations (11.33) and (11.34)) become:

$$X_1 + X_2 + X_3 = 0 \quad (11.35)$$

$$\frac{X_1}{X_2} = \beta \quad (11.36)$$

For a FET,  $R_i = \infty$  and  $R_o = r_d$ . Hence, the conditions of oscillation (equations (11.33) and (11.34)) become:

$$X_1 + X_2 + X_3 = 0 \quad (11.37)$$

$$\frac{X_1}{X_2} = g_m r_d \quad (11.38)$$

According to equations 11.36 and 11.38,  $X_1$  and  $X_2$  must be of the same sign. Furthermore, according to equations 11.35 and 11.37,  $X_3$  must be of the opposite sign. There are two possibilities, first  $X_1$  and  $X_2$  have negative reactance (capacitors), while  $X_3$  has a positive reactance (inductor). This type of oscillator is called a Colpitts oscillator. The second possibility is that  $X_1$  and  $X_2$  have positive reactance (inductors), while  $X_3$  has a negative reactance (capacitor). This type of oscillator is called a Hartley oscillator.

Figure 11.24 shows a circuit diagram of a Colpitts oscillator [108]. The frequency of oscillation can be found by substituting the actual reactance in equation 11.35. This leads to a frequency of oscillation given by:

$$f_{\text{osc}} = \frac{1}{2\pi} \sqrt{\frac{C_1 + C_2}{LC_1 C_2}} \quad (11.39)$$

Figure 11.25 shows a circuit diagram of a Hartley oscillator [109]. The frequency of oscillation can be found by substituting the actual reactance in equation 11.35. This leads to a frequency of oscillation given by:

$$f_{\text{osc}} = \frac{1}{2\pi} \frac{1}{\sqrt{C(L_1 + L_2)}} \quad (11.40)$$

### 11.5.3 The Negative Resistance Approach

Oscillators are considered to be negative resistance devices [110]. The positive resistance (normal resistance) dissipates power when an electrical current passes through it. The negative resistance is the opposite of this, it generates power when a current passes through it. For a positive resistance, the direction of flow of the current is that of the voltage drop. For a negative resistance the direction of flow of the current is that of a voltage rise, this is what makes the resistance negative.

In an oscillator, the power generated by the oscillator's [equivalent] negative resistance equals the power dissipated in the lossy parts of the circuit. This is the equilibrium condition that sustains oscillation.

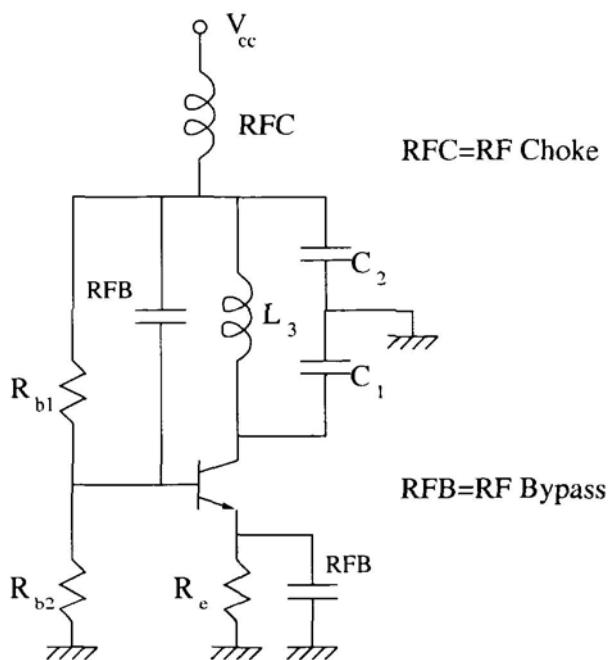


Figure 11.24. Circuit diagram of a Colpitts Oscillator.

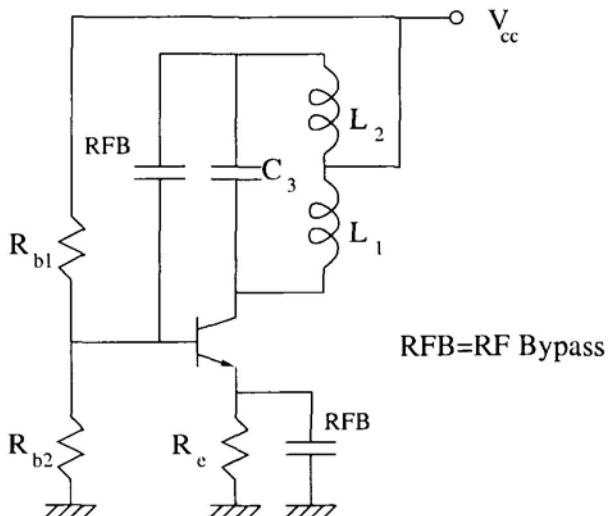


Figure 11.25. Circuit diagram of a Hartley Oscillator.

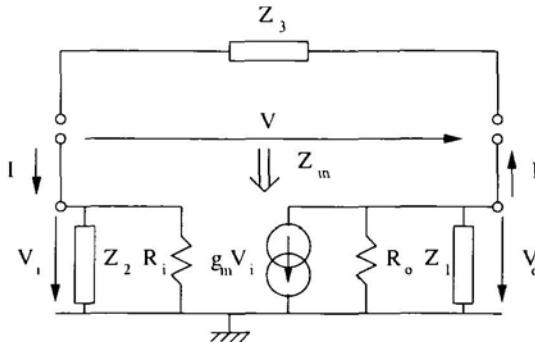


Figure 11.26. Equivalent circuit of the Colpitts/Hartley oscillator illustrating the oscillator's negative resistance.

The equivalent circuit of the Colpitts/Hartley oscillator shown in Figure 11.23 has been redrawn in Figure 11.26. The input impedance  $Z_{in}$  of the oscillator, between the two terminals  $Z_3$  is connected to, is given by:

$$Z_{in} = [Z_2 \parallel R_i] + \{1 + g_m[Z_2 \parallel R_i]\}[R_o \parallel Z_1] \quad (11.41)$$

The condition of oscillation is that the effect of  $Z_3$  is canceled by  $Z_{in}$ . Therefore, for oscillations:

$$Z_{in} + Z_3 = 0 \quad (11.42)$$

Combining equations (11.41) and 11.42 we get the same condition of oscillation previously obtained in equation (11.29). To gain more insight into the negative resistance concept, we will consider the case of a FET oscillator in detail. For a FET,  $R_{in} = \infty$  and  $R_o = r_d$ . Hence, the input impedance of a FET oscillator from equation (11.41) becomes:

$$Z_{in} = Z_2 + [1 + g_m Z_2][r_d \parallel Z_1] \quad (11.43)$$

Taking  $Z_1 = jX_1$  and  $Z_2 = jX_2$ , equation (11.43) becomes:

$$Z_{in} = \frac{X_1^2 R_o - X_1 X_2 R_o^2 g_m}{X_1^2 + R_o^2} + j \left( X_2 + \frac{X_1 R_o^2 + X_1^2 X_2 R_o g_m}{X_1^2 + R_o^2} \right) \quad (11.44)$$

If  $X_1 X_2 R_o^2 g_m > X_1^2 R_o$  the real part of  $Z_{in}$  is negative. Ideally,  $Z_3$  is a pure reactance. Practically, however,  $Z_3$  will have some losses, which make it

have a positive resistive component. This positive resistance is canceled out by the negative resistance of  $Z_{in}$ . In this sense, the negative resistance of the oscillator is used to compensate the energy dissipated in the lossy components.

A variant of the Colpitts oscillator is the Clapp-Gouriet Oscillator [111]. The Clapp oscillator has an extra capacitor  $C_3$  in series with the inductor. This oscillator is tuned by changing a single capacitor  $C_3$ . At the resonance frequency, the series combination of  $L$  and  $C_3$  has an inductive reactance that changes with frequency at a greater rate than the reactance of a single inductor. This makes the Clapp oscillator more frequency stable, than the Colpitts oscillator, provided that the inductor is stable [106]. Figure 11.27 shows the circuit diagram of the Clapp oscillator. The frequency of oscillation is given by:

$$f_{osc} = \frac{1}{2} \sqrt{\frac{1}{L} \left( \frac{1}{C_1} + \frac{1}{C_2} + \frac{1}{C_3} \right)} \quad (11.45)$$

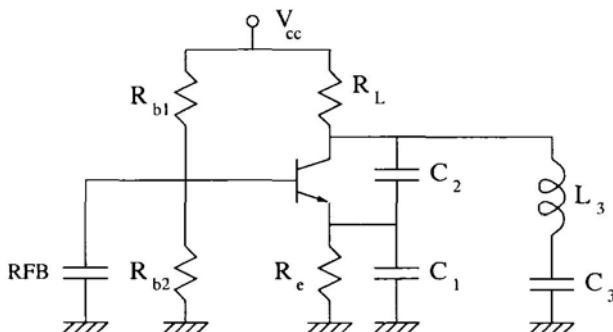


Figure 11.27. Circuit diagram of the Clapp-Gouriet Oscillator.

### 11.5.4 Crystal Oscillators

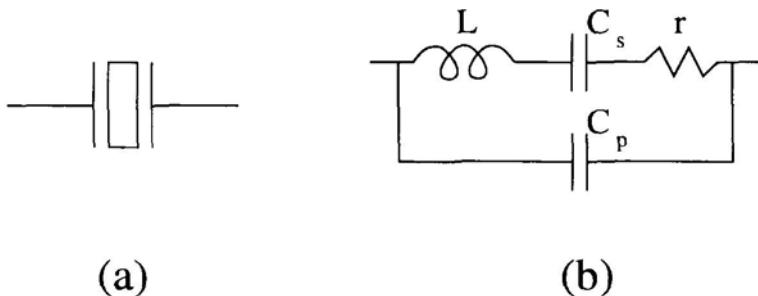
Crystal oscillators rely on the piezo-electric effect of crystalline material, which was discovered in the late 19th century by Pierre and Jacques Currie. During the 1920's and 1930's crystal oscillators started to find their way into radio applications.

The crystal oscillator consists of a piezo-electric crystal, such as a quartz, connected between two electrodes. The applied electrical energy is converted into vibratory mechanical energy, which is then re-converted back into electrical energy at the terminals of the resonator [112].

The crystal oscillators are characterized by their high quality factor, which can be several hundred thousand, this makes the crystal oscillator have short term frequency stability. Crystal oscillators are also characterized by their very

stable electromechanical characteristics with temperature and time (a few 10's ppm (part-per-million)), this leads to long term frequency stability. Crystal oscillators can have resonant frequencies varying from the KHz range to the sub-GHz range.

The electrical symbol for a crystal oscillator is shown in Figure 11.28.a. The equivalent circuit is shown in Figure 11.28.b. A more detailed equivalent circuit can be found in [112]. The parallel capacitor  $C_p$  is much larger than the series one  $C_s$ . Their ratio can vary from 100 to 20000 [112]. The resistance  $r$  determines the quality factor,  $Q = wL/r$ . The inductor L can be in the hundred of Henries range [108], but usually it's in the 10's of mh range.



*Figure 11.28. (a) Circuit symbol of a crystal oscillator. (b) Equivalent circuit of the crystal oscillator.*

The circuit in Figure 11.28.b has two resonant frequencies. A series resonant frequency and a parallel resonant frequency. At the series resonance frequency, the impedance approaches zero (assuming  $r = 0$ ). The series resonant frequency is given by:

$$f_s = \frac{1}{2\pi} \frac{1}{\sqrt{LC_s}} \quad (11.46)$$

The parallel resonant frequency, also known as anti-resonant frequency is the frequency at which the impedance approaches  $\infty$  (assuming  $r = 0$ ).

$$f_p = \frac{1}{2\pi} \frac{1}{\sqrt{L \frac{C_s C_p}{C_s + C_p}}} \quad (11.47)$$

The crystal reactance varies with frequency as shown in Figure 11.29. Between the series resonant frequency and the parallel resonant frequency the impedance is inductive, and changes at fast rate with frequency. In an oscillator circuit this is the region that the crystal should operate in. Hence, the crystal oscillator replaces the inductor in the oscillator circuit. Notice that, the rapid change of inductance with frequency helps stabilize the frequency of the crystal oscillator.

One form of a crystal oscillator circuit is a Colpitts oscillator with the inductor replaced by the crystal. This type of oscillator is known as the Pierce oscillator, and is shown in Figure 11.30.

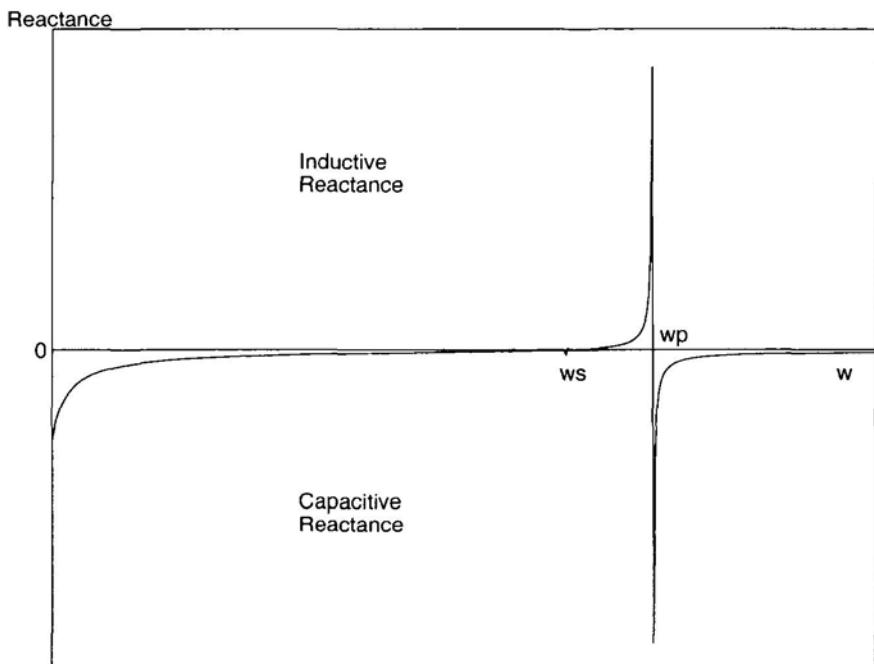


Figure 11.29. Crystal oscillator reactance versus frequency.

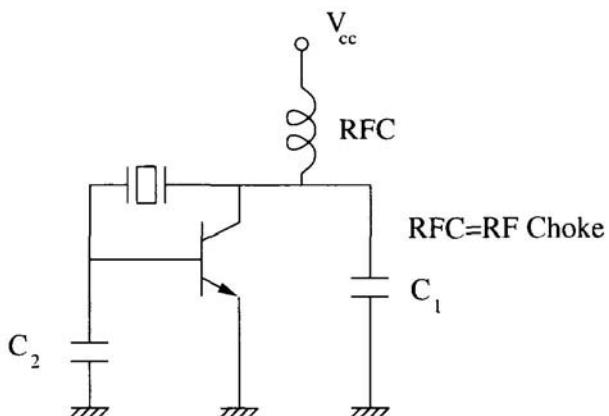


Figure 11.30. Circuit diagram of the Pierce Oscillator, the bias circuit is not shown.

The main disadvantage of crystal oscillator circuits is that they are fixed frequency oscillators. To change the frequency of oscillation it is possible to add a capacitor in shunt with the crystal oscillator this makes the parallel resonant frequency move towards the series resonant frequency, but even then the tuning range is limited. The difference between the series resonant and the parallel resonant frequency is called the pulling range and this is given by:

$$f_{pulling} = \frac{C_s}{2C_p} f_s \quad (11.48)$$

### 11.5.5 Voltage Controlled Oscillators

The voltage-controlled oscillator (VCO) is a device that produces an output signal whose frequency of oscillation is dependent on its input voltage. It is desired to keep the relation between the output frequency and the input voltage a linear relation as given by the following equation:

Where,  $f_{osc} = f_o + K'_o v_{in}$  (11.49)

$f_o$  is the frequency of the output signal of the VCO when  $v_{in} = 0$ .

$v_{in}$  is the voltage of the input signal.

$K'_o$  is the VCO constant in Hz/Volt.

A voltage controlled oscillator uses a voltage dependant reactance element, such as a varactor, to control the output frequency of the VCO.

## Chapter 12

# FREQUENCY SYNTHESIZERS

### 12.1 INTRODUCTION

A frequency synthesizer is defined as an electronic device, which is capable of producing one or many frequencies from one or more reference sources [113]. Generally, the output frequencies are any fraction of the input reference, as shown in Figure 12.1. Frequency synthesizers today have a wide variety of applications including radios, satellites, radar systems, television receivers, test equipment, cellular phones, and personal communications services (PCS).

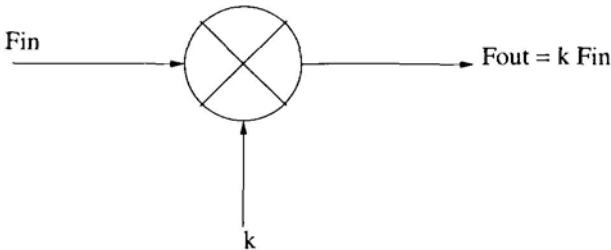


Figure 12.1. General frequency synthesizer structure.

The recent increased demand in portable services, such as cellular phones and PCS, has dramatically increased the use of available spectrum. The main component in such systems, which enables channel selection is the frequency synthesizer. The accuracy of selecting a channel with the purest possible reference source has been subject to intense research in the past few years. Its simple structure combines three large disciplines: analog and digital circuit techniques, digital signal processing, and control theory.

In section 12.2, we present several important frequency synthesizer parameters. In section 12.3, we present two frequency synthesis techniques: the

fractional-N frequency synthesizer and the direct digital frequency synthesizer. In section 12.4, we discuss the phase noise in frequency synthesizers. Finally, we conclude the chapter with a summary in section 12.5.

## 12.2 FREQUENCY SYNTHESIZER (FS) PARAMETERS

There are several frequency synthesizer parameters that are important in both design and analysis. In this section, the important FS parameters are listed.

### 12.2.1 Frequency Range

One important parameter of a frequency synthesizer is its frequency range. This is defined by the nature of the application. Two important parameters, which control the frequency range, are the number of required channels and the channel spacing. For mobile commercial applications this may range from 900MHz (ISM Band) to 2.5GHz (PCS). For radar and satellite applications, the operating frequency may be extended to several GHz.

### 12.2.2 Frequency Step Size

The frequency step size dictates the resolution requirement of the frequency synthesizer. If very narrow frequency step sizes are required, then a very sharp frequency synthesizer is required. Such a requirement may result from wireless standards with small channel bandwidths (such as the Mobitex wireless data communications standard).

### 12.2.3 Output Power Level

The output power level is the strength of the output signal of the frequency synthesizer. This parameter is usually reported in dBm. The dBm scale is normalized to one 1 mW.

### 12.2.4 Switching Speed

The switching speed of a frequency synthesizer is defined to be the time interval necessary to switch from one frequency to another. This parameter is only important in applications such as frequency hopping spread spectrum techniques, where more than one frequency is used in the communications channel, and in half-duplex systems with different transmit-receive frequencies. This parameter depends on the frequency synthesizer architecture.

### 12.2.5 Phase Noise

One of the most important parameters in frequency synthesizers is the phase noise. It is important to make a distinction between additive noise and phase noise. A non-ideal sine wave may be given as:

$$V(t) = V[1 + n_1(t)] \sin[(\omega_0 t + n_2(t))] \quad (12.1)$$

Consider a sine wave, due to thermal noise in the electronic devices in a frequency synthesizer, the sine wave may be corrupted by both amplitude noise (due to amplitude modulation (AM)) and phase noise. In the case of AM noise, only the amplitude is affected (i.e.  $n_1(t)$  represents AM noise in the above equation); whereas in the case of phase noise, horizontal excursions superimposed on the sine wave exist (i.e. this is represented by  $n_2(t)$  in the above equation). This results in an error term associated with each period of the synthesized sine wave. For example, if a period  $P = 5$  nsec is required, at a certain instance a period of 5.1 ns may be produced; whereas in another instance a period of 4.8 ns may be produced. A sine wave in the frequency domain is represented by two pulses in positive and negative frequencies, as shown in Figure 12.2.a. A sine wave corrupted by phase noise results in smearing of this sine wave as shown in Figure 12.2.b.

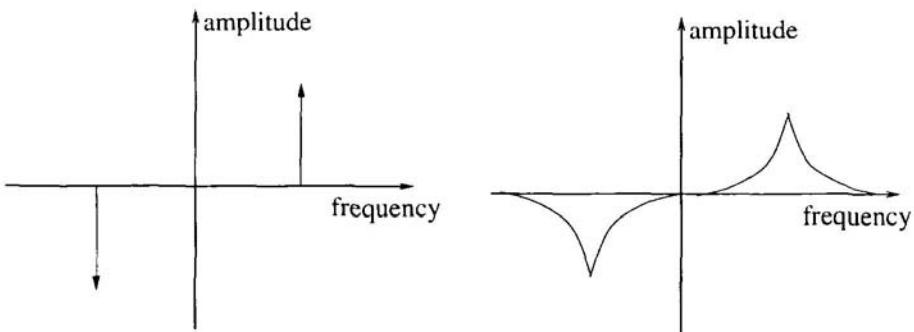
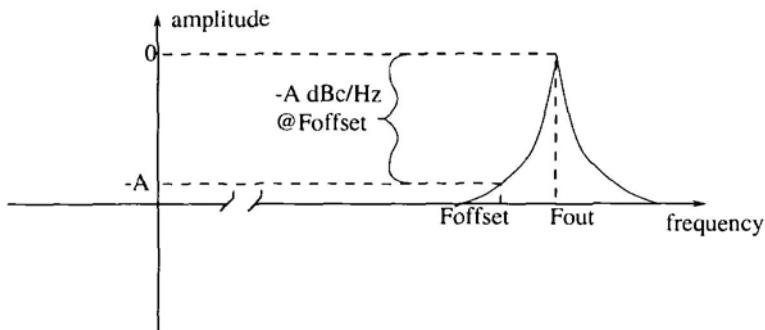


Figure 12.2. Frequency representation of a sine wave (a) in ideal form and (b) corrupted by phase noise.

Phase noise is usually reported as the difference in magnitude between the carrier power and the sideband power in 1-Hz bandwidth at a specified frequency offset. The units of phase noise are dBc/Hz @ frequency offset. In communication systems, the phase noise requirement in the system depends on the choice of modulation scheme (FSK, MSK, QPSK, QAM) as well as the expected input signal levels and adjacent channel levels. Figure 12.3 shows a typical output spectrum of a frequency synthesizer.



*Figure 12.3.* Typical output spectrum of a frequency synthesizer.

### 12.2.6 Spurious Output

In addition to phase noise, spurious signals due to non-linearities in the electronics of the frequency synthesizer cause spurious outputs. As stated earlier, ideally the single-sided output spectrum of a frequency synthesizer should be a single pulse at the desired output frequency. Non-linearities in the electronic devices generate harmonics, which are intermodulated together. It is this intermodulation that may cause otherwise far removed harmonic signals to enter back into frequency ranges close to the carrier frequency.

### 12.2.7 Power Consumption

One important parameter in frequency synthesizers is the power consumption. This is especially important in portable applications, such as cellular phones and laptops equipped with wireless modems. Minimizing this parameter usually comes at the expense of increased phase noise.

## 12.3 FREQUENCY SYNTHESIZER TECHNIQUES

There are two main methods of synthesizing sine/cosine functions in use today. The most popular method of frequency synthesis is based on a feedback mechanism in which the reference frequency is multiplied to a higher frequency. This method is called phase-locked loop (PLL). The reason for its popularity stems from its feasibility to implement such a device in monolithic form (single chip) at a low cost. This has been made possible by advances in silicon technology.

The other method of frequency synthesis is called direct digital frequency synthesis (DDFS). The DDFS architecture is based on a purely digital solution employing digital signal processing techniques. In this synthesis technique, a digital frequency word is given as input and a pure digital sine/cosine

waveform appears at the output. This type of frequency synthesis is less popular in front-end wireless systems due to its limited output frequency.

### 12.3.1 Fractional-N PLL Synthesis

One of the most common methods of synthesizing frequencies for RF wireless applications is through the use of charge pump phase-locked loops (PLLs). PLL synthesis is an indirect frequency synthesis approach by which a low frequency, high spectral quality signal is multiplied to a higher frequency with minimum addition of phase noise and spurs. Phase locked loops were presented in chapter 11.

Although simple, the performance of the single-loop PLL is usually insufficient for RF applications. Consider, for example, the Mobitex wireless standard which requires 12.5KHz channel spacing [114]. This restricts the reference frequency to 12.5KHz (output of the PLL is always an integer ratio of the reference frequency). If an output frequency of 896MHz is desired, the division ratio must be 71,680. Since the phase noise contribution of the frequency divider is  $20 \log_{10}(N)$  [1], the output signal would be degraded by as much as 97dB, which is unacceptable by most wireless communication standards.

Fractional-N PLLs offer higher spectral purity than single loop PLLs. As shown in Figure 12.4, frequency spacing smaller than the input frequency is achieved by selecting between 2 division ratios of either  $N$  or  $N + 1$  [114], [115]. The selection is done in such a way that the average division ratio over several cycles is  $\{Nf + (N + 1)(1 - f)\}$ , where  $N$  is the integer division ratio and  $f$  is a fraction between 0 and 1. For example, if a division ratio of 4 is chosen for four clock cycles and a division ratio of 5 is chosen for the fifth clock cycle, an average division ratio of  $(4 * 4 + 5 * 1)/5 = 4.2$  is achieved.

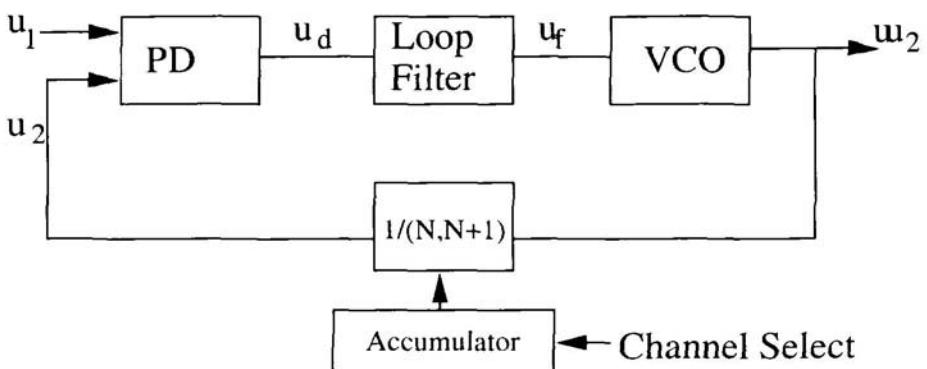


Figure 12.4. Fractional-N PLL architecture.

The operation of the fractional-N PLL is best described through an example. Suppose that a frequency channel spacing of 30 kHz is desired and the desired output frequency range is from 900 MHz to 910 MHz. For a conventional PLL, the minimum division ratio would have to be  $N = 900\text{MHz}/30\text{kHz} = 30,000$ . If a fractional-N PLL can have fractional ratios of [1/5, 2/5, 3/5, 4/5, 5/5], then the necessary division ratio would be  $N = 900\text{MHz}/(5 * 30\text{kHz}) = 6000$ . These fraction ratios listed above may be realized by a modulo 5 counter, where an overflow signal would indicate jumping from a ratio of  $N$  to  $N + 1$  for a period equal to that of the reference clock. In general, to achieve a frequency resolution smaller than the input frequency by a factor of  $M$ , a modulo  $M$  counter is needed ( $M$  in the above example is 5). A summary of calculations for both cases (conventional and fractional-N PLL) are shown in Table 12.1 below.

*Table 12.1. Conventional versus Fractional-N Frequency Dividers.*

Conventional PLL	Fractional-N PLL
$900.00 = 0.03(30,000)\text{MHz}$	$900.00 = f_{comp}(6000 + 0/5)\text{MHz}$
$900.03 = 0.03(30,001)\text{MHz}$	$900.03 = f_{comp}(6000 + 1/5)\text{MHz}$
$900.06 = 0.03(30,002)\text{MHz}$	$900.06 = f_{comp}(6000 + 2/5)\text{MHz}$

$f_{comp}$  is the phase detector operation frequency =  $5 * (\text{channel spacing}) = 0.15\text{MHz}$ . Table 12.1 demonstrates an important feature of fractional-N PLLs over conventional PLLs, which is the lower frequency division ratio. This is achieved by allowing the reference frequency, which in this case is equal to  $f_{comp}$ , to be larger than the channel spacing. There are several advantages for the lower division ratio such as reduced hardware complexity, lower phase noise due to the logarithmic dependence of the noise on the division ratio [114], and improved switching time (due to the larger loop filter bandwidth).

The fractional-N PLL architecture, however, does come with its disadvantages. First, the output frequency range is very narrow. This is because the division ratio differs only by a factor of one. This could be solved by increasing the number of division ratios. This solution, however, comes at the expense of additional hardware complexity.

A generic fractional-N frequency divider is shown in Figure 12.5. It consists of a prescaler, low speed asynchronous divider, and a phase error compensator. The prescaler is the component of the divider which receives its input directly from the VCO. Frequency dividers are usually implemented as counters. The logic of the prescaler is usually implemented using a modified Johnson counter (for a discussion on Johnson counters refer to [116]). A logic level diagram of a divide by 4/5 prescaler is shown in Figure 12.6. The prescaler is designed in such a way that if  $MC = 0$ , the frequency division ratio is 4, and if  $MC = 1$ , the frequency division ratio is five. The purpose of this device is to effectively

prevent an entire VCO period from being fed back to the phase detector. The output of the prescaler is then fed into an asynchronous divider. The asynchronous divider consists of nothing but cascaded divide by 2 circuits, usually implemented as T flip-flops.

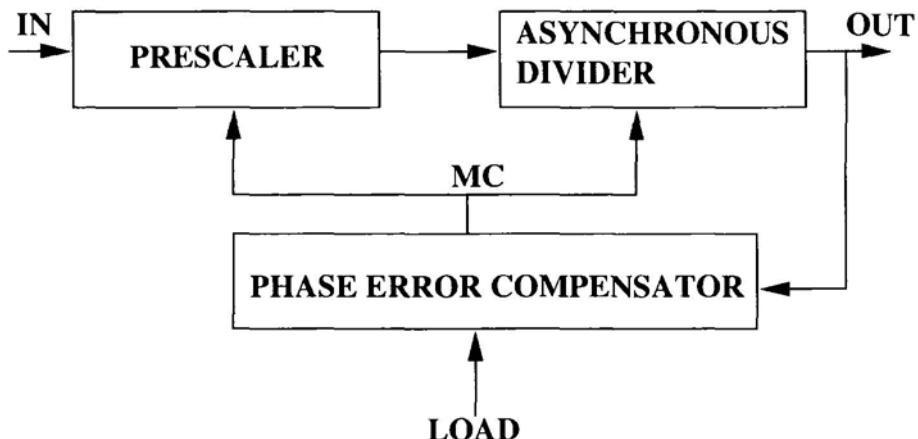


Figure 12.5. Block diagram of a generic fractional-N frequency divider.

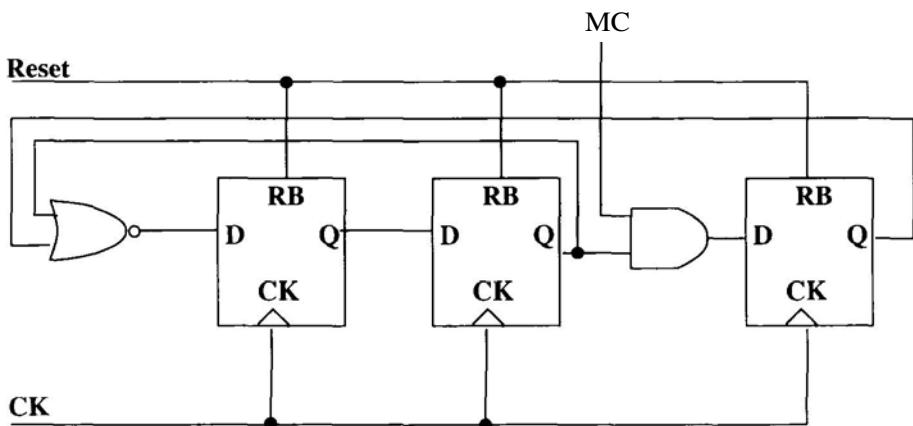


Figure 12.6. Divide-by-4/5 prescaler.

The phase error compensator ensures that a steady fractional division ratio is given as specified by the LOAD input shown in Figure 12.5. The conventional method of realizing the phase error compensator is by implementing it as an accumulator. The input to the accumulator would be the desired fractional ratio. The overflow output would be used to change the division ratio from  $N$  to  $N + 1$ . Such an implementation is called a pulse swallower counter. The

width (number of bits) of the counter is equal to  $\log_2$  [reference frequency divided by the minimum frequency resolution required]. Suppose, for example, that the PLL's reference frequency ( $F_{ref}$ ) is 7MHz and the desired channel spacing of the frequency synthesizer is 12.5 KHz. The required number of bits of the accumulator is  $\lceil \log_2(560) \rceil = 10$  bits. A typical waveform of the value stored in the accumulator of a pulse swallower counter is shown in Figure 12.7. The average value of this sawtooth function should be the required fractional ratio. The undesired ac component of this signal causes modulation of the VCO output frequency, and results in spurs in the output spectrum. The rising edge of each segment of the sawtooth waveform represents the phase detector output for that period. Figure 12.8 shows a typical frequency spectrum of a pulse swallower based fractional-N divider. Note that the maximum strength is centered around the desired fractional value. The other frequency components are undesired spurs.

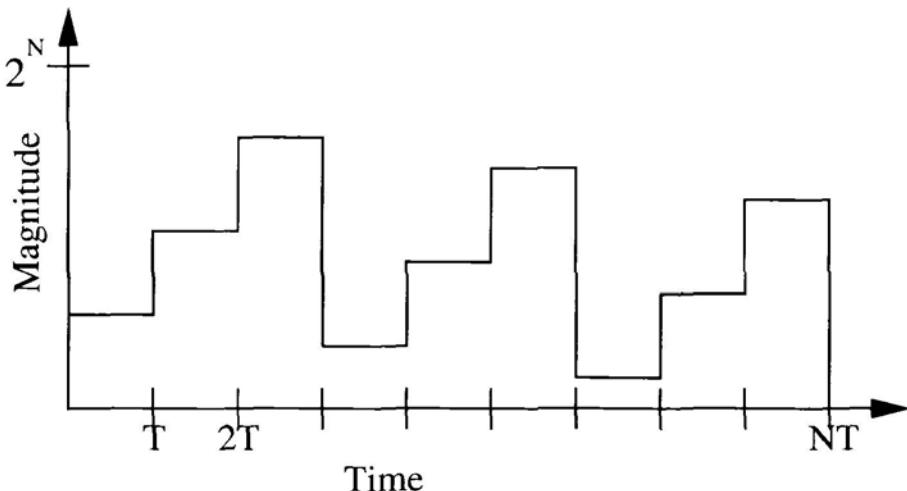


Figure 12.7. Typical values stored in an accumulator of a pulse swallower counter.

An important issue to understand about fractional-N synthesizers is the spurious response. Thermal noise and other circuit non-linearities contribute to spurs in the output spectrum of frequency synthesizers. In fractional-N synthesizers, additional sources of spurs may originate from changing the feedback division ratio from one value to the other. As stated earlier, a pulse swallower counter causes undesired spurs due to the ac component shown in Figure 12.7. If the closed loop bandwidth of the PLL is smaller than the frequency of the spurs, then the closed loop filter of the PLL itself may be used to effectively inhibit the spurs. However, most practical frequency synthesizers require very narrow channel spacing. This means that the spurs would be very

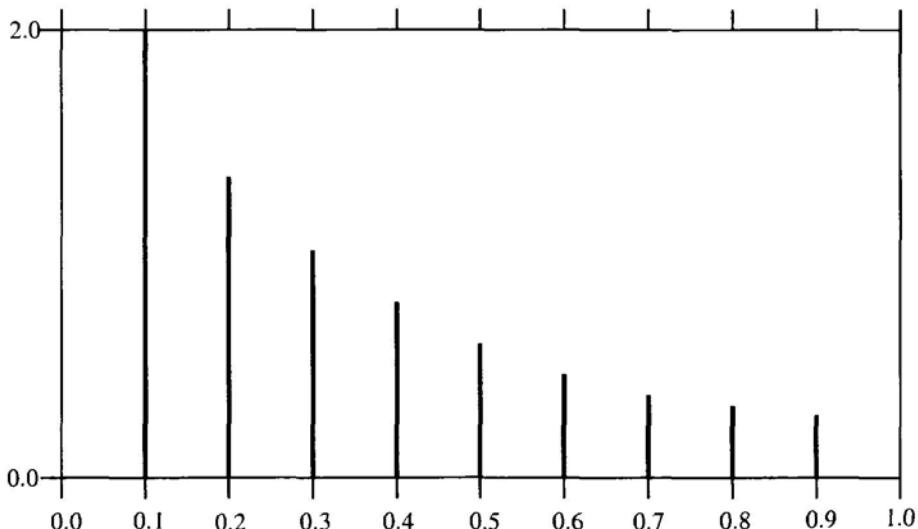


Figure 12.8. Frequency spectrum of spurs in a fractional-N divider [1].

close to the carrier frequency. This would entail making the closed loop bandwidth very narrow, which may cause excessively long lock times as well as insufficient filtration of VCO phase noise near the carrier frequency.

One solution that has been proposed is the use of analog phase interpolation (API) circuitry as shown in Figure 12.9. As shown in Figure 12.7, the value stored in the pulse swallower counter may be greater than or less than the required fractional value. The greater the value, the larger the phase error in the phase detector. This variation in the error in the phase detector causes fluctuations in the VCO's output frequency, and hence spurs are generated. In order to eliminate these spurs, the value residing in the pulse swallower counter may be applied to a digital-to-analog converter, which would be summed to the control voltage of the VCO. This, in effect would cancel out the variation of the phase error in the phase detector. It has been reported that this method may cancel spurs only down to -60 dBc. The reason for this limitation is due to circuit non-linearities associated with the digital-to-analog converter [1].

More recently, pure digital techniques have been proposed to eliminate spurs in fractional-N synthesizers. One such technique rely on noise shaping techniques, such as those usually associated with sigma-delta ( $\Sigma - \Delta$ ) converters. One such technique explicitly uses a digital version of the  $\Sigma - \Delta$  modulations scheme [117], [118].  $\Sigma - \Delta$  modulators have been studied extensively in the context of oversampled analog-to-digital converters, see section 7.7 of chapter 7. In oversampled analog-to-digital converters, a 1-b quantizer is used. This produces large quantization noise. This quantization noise is filtered out

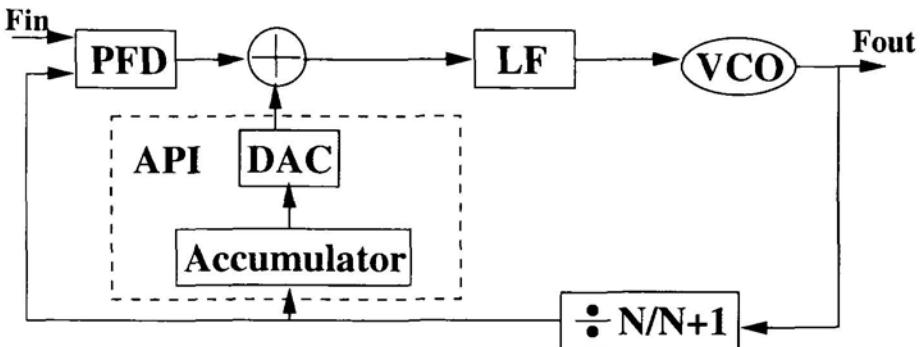


Figure 12.9. Block diagram of API used in a fractional-N PLL synthesizer.

by the noise shaping characteristics of the  $\Sigma - \Delta$  modulator. The  $\Sigma - \Delta$  modulator is used to shift the quantization noise to a high frequency while still preserving the spectrum of the input signal. The high frequency noise would then be filtered by a low-pass filter. A similar technique may be used in fractional-N PLL frequency synthesizers, since the closed loop bandwidth is typically a low value, and hence the closed loop would act as a low-pass filter. The quantization noise in fractional-N PLLs would be the frequency spurs due to frequency division ratio switching. Hence, the  $\Sigma - \Delta$  modulator transfers correlated frequency spurs into high frequency random noise. A block level diagram of a digital version of the  $\Sigma - \Delta$  modulator is shown in Figure 12.10.

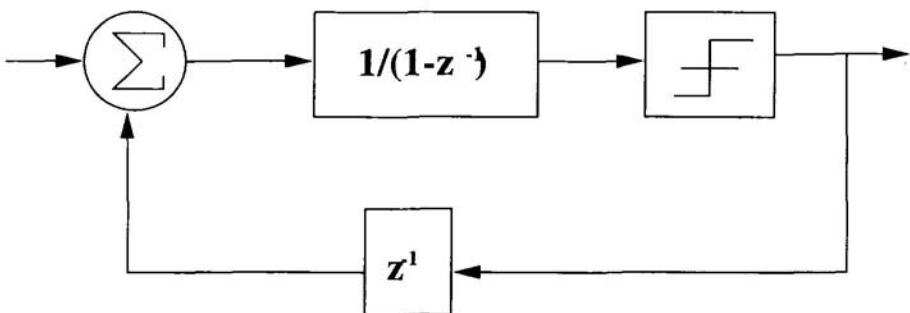


Figure 12.10. Digital Sigma-Delta modulator.

Digital  $\Sigma - \Delta$  modulators differ from the pulse swallowing (PS) approach in two ways:

- PS input range is from 0 to  $2k$ , where  $k$  is the width of the counter. In  $\Sigma - \Delta$  modulators, the input range is  $\pm M$ , where  $M = 2k - 1$ .

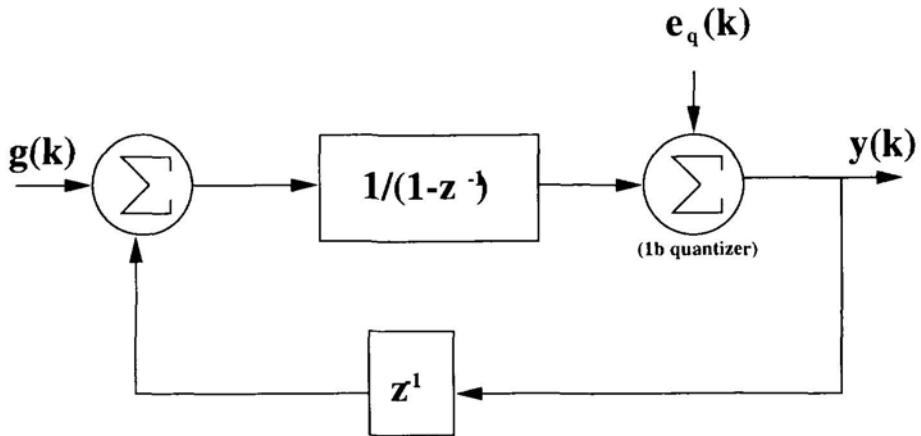


Figure 12.11. Conventional first-order Sigma-Delta modulator.

- PS produces 1's and 0's centered around a digital value of  $2^{k-1}$ ; where as in  $\Sigma - \Delta$  modulators, a stream of 1's and 0's are produced centered around a digital value of 0.

The frequency shaping property of  $\Sigma - \Delta$  modulators is best understood in terms of its transfer function. A diagram of a conventional first order  $\Sigma - \Delta$  modulator is shown in Figure 12.11. Note that, the quantization noise is modeled as additive noise. The justification of this assumption is beyond the scope of this book; however, the interested reader may refer to [119]. The transfer function of this feedback loop is:

$$Y(z) = \frac{1/(1-z^{-1})}{1 + \frac{z^{-1}}{1-z^{-1}}} G(z) + \frac{E_q(z)}{1 + \frac{z^{-1}}{1-z^{-1}}} \quad (12.2)$$

which may be simplified to

$$Y(z) = G(z) + (1 - z^{-1}) E_q(z) \quad (12.3)$$

In general, for nth order  $\Sigma - \Delta$  modulators, the output of the modulator becomes [118]:

$$Y(z) = G(z) + (1 - z^{-1}) E_{qn}(z) \quad (12.4)$$

where,  $E_{qn}$  is the quantization noise of the nth order sigma-delta modulator. In the context of the fractional-N PLLs, the output frequency becomes [118]:

$$f_{out}(z) = N \cdot G(z) \times F_{ref} + (1 - z^{-1})^n F_{ref} E_{qn}(z) \quad (12.5)$$

The power spectral density of the quantization noise, in the continuous frequency domain is given by:

$$\xi(f) = \frac{(2\pi)^2}{12F_{ref}} \left[ \frac{f}{F_{ref}/2\pi} \right]^{2(m-1)} \quad (12.6)$$

where,  $m$  is the order of the modulator and  $f$  is the operating frequency. The disadvantage of increasing the size of the modulator is the increase in hardware complexity. Note that the quantization noise rises at 20 dB per decade using a first order  $\Sigma - \Delta$  modulator, and 40 dB per decade using a second order  $\Sigma - \Delta$  modulator. This rise imposes an upper limit on the order of the  $\Sigma - \Delta$  modulator beyond which the frequency spurs become too hard to filter out by the PLL. Figure 12.12 shows the effect of a  $\Sigma - \Delta$  modulator on the spurs generated by the frequency divider in a closed loop PLL. For a first order  $\Sigma - \Delta$  modulator, the frequency spurs are filtered to less than -100 dB/Hz for all offset frequencies. This is more or less the same level of spurs as that produced by a regular fractional-N frequency synthesizer. Note that, the second order  $\Sigma - \Delta$  modulator has less spurs at low frequency offsets, but it has less spur suppression capabilities at higher frequency offsets due to the sharp rise of the quantization noise with frequency offset.

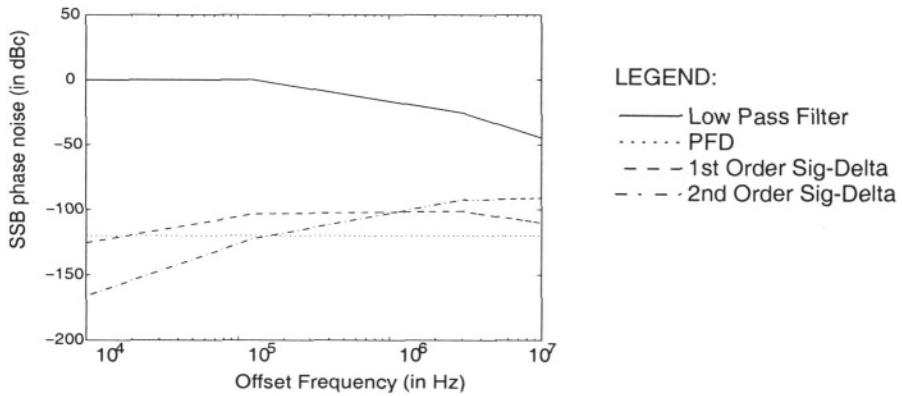


Figure 12.12. Effect of Sigma-Delta modulator on spurs of fractional-N synthesizers [2].

In the digital version of a  $\Sigma - \Delta$  modulator, the input frequency word (fractional part which is fed into the phase error compensator), must be integrated to produce a phase, since  $\Sigma - \Delta$  modulators operate on reducing the phase noise. In the digital domain an integrator may be realized by an accumulator. Note however, that an accumulator would add a pole, or a factor of  $1/(1 - z - 1)$  in the digital domain, to the expression in equation 12.4. This has the effect of reducing the order of the  $\Sigma - \Delta$  modulator by one. Therefore,

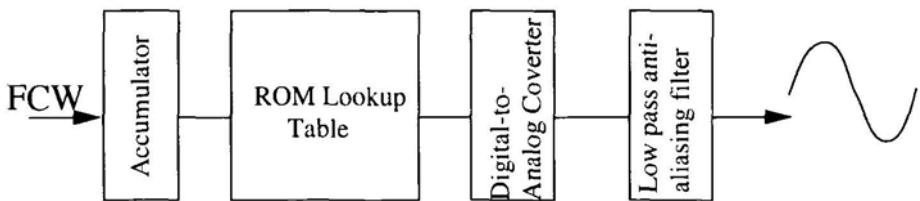


Figure 12.13. Direct digital frequency synthesizer architecture.

a second order digital  $\Sigma - \Delta$  modulator must be realized in order to obtain the same noise shaping characteristics as a conventional first order  $\Sigma - \Delta$  modulator.

### 12.3.2 Direct Digital Synthesis

As stated earlier, direct digital frequency synthesizers present a fully digital solution to waveform synthesis. A block diagram of a DDFS is shown in Figure 12.13. The operation of the DDFS is as follows. The frequency control word (FCW) is fed into an accumulator (of length  $j$  bits), which acts as a digital integrator, to produce a phase word. The phase word acts as an address to the ROM look up table, which contains the values of the function for the given value of phase. In order to reduce the size of the ROM table, only  $k < j$  bits are used to address the ROM table. The digital output is then fed into a digital-to-analog converter (DAC) to produce the desired analog signal (which is usually a sine or cosine function in the case of frequency synthesis). A low pass anti-aliasing filter is also needed after the DAC. The output frequency is given as:

$$f_{out} = \frac{FCW f_{clk}}{2^j} \quad (12.7)$$

where, FCW is the frequency control word,  $f_{clk}$  is the frequency of the clock signal, and  $j$  is the number of accumulator bits. In order to satisfy the Nyquist criterion, the maximum output frequency is given as:

$$f_{out} = \frac{f_{clk}}{2} \quad (12.8)$$

There are four advantages to using a DDFS as opposed to a PLL. Firstly, DDFS is a pure digital solution (except for the DAC), which means that its performance is easily and reliably repeatable and it can be easily integrated with other on-chip circuitry. Another advantage is its inherent instantaneous lock time. Such a characteristic is important for fast frequency hopping spread spectrum (FHSS) applications. The third advantage is that it is the only known

frequency synthesis technique which allows linear phase shifting (important for PSK modulation scheme) [114]. The fourth advantage is that, unlike PLLs, no phase noise is added to the reference frequency.

The main drawback of the DDFS architecture is the high level of spurious frequencies. This is primarily generated by the finite word length representation of phase and amplitude as well as non-idealities in the digital-to-analog converter (DAC). Although there are several proposed techniques to reduce the level and number of spurs generated by the DDFS architecture [120], [121], [122], [123], [124] and [125], it is still regarded to be inferior to that of the PLL architecture.

Another drawback of the DDFS architecture is the limited output frequency. As stated earlier, the maximum output frequency is half the clock frequency. If, for example, a 1 GHz sine wave is required, the digital circuitry must be able to operate at 2 GHz. This is a very stringent requirement, even with modern CMOS technologies. Another disadvantage is that high frequency DACs are very difficult to design, and are now regarded as being the limiting factor in DDFS performance [114].

### **12.3.3 Analysis and Reduction of Spurs in Direct Digital Synthesis**

As mentioned earlier, one of the main drawbacks of the DDFS technique is its spurious signal performance. In this section, the sources of such spurs are first analyzed. This is followed by a quick review of current techniques used to limit spurs in DDFS.

In general, spurs in DDFS are generated by the periodicity of the quantization errors due to finite word length effects in various components in the DDFS. There are three major contributors of spurs in the DDFS architecture. The rate of overflows in the phase accumulator determines the output frequency of the DDFS, which is given by equation (12.7). Note that, if the frequency control word (FCW) does not divide the length of the accumulator  $2j$ , this would create uneven instantaneous period lengths. This irregularity of period lengths is the contributor to spur source 1. It has been shown in [121] that the carrier-to-spur rate is given by:

$$\left(\frac{C}{S}\right)_1 = (6.02k - 3.992)\text{dBc} \quad (12.9)$$

where,  $k$  is the word length of the phase accumulator used to address to ROM lookup table. This indicates a 6 dB reduction of spur strength for each address bit to the ROM; i.e. 6 dB improvement is at the expense of doubling the size of the ROM lookup table.

The second source of spurs is generated by finite word length effects in the ROM lookup table. As stated earlier, the ROM lookup table contains the value

of the periodic function to be synthesized, usually a sine or cosine function. Simulations have indicated that the majority of spurs created by finite word length effects in the ROM lookup table are very small and can be treated as background noise [123]. There are, however, a few spurs (namely 3rd order harmonic spurs) that have a noticeably high spur level and their strength is estimated by:

$$\left(\frac{C}{S}\right)_2 = (6S + 14)\text{dBc} \quad (12.10)$$

where,  $S$  is the number of bits used to represent the amplitude of the periodic function to be synthesized.

Spur source 3 is created by non-linearities in the digital-to-analog converter (DAC). This source of spur is generally the most dominant and usually limits DDFS performance. A closed loop expression for this source of spur is difficult to obtain since it depends on the underlying electronic circuitry's performance measurements such as accuracy, dynamic range, intermodulation products, linearity, and dynamic characteristics [114].

## 12.4 ANALYZING PHASE NOISE IN FREQUENCY SYNTHESIZERS

One important frequency synthesizer parameter describing the quality of the synthesized frequency is phase noise. As stated earlier, ideally the spectrum of the frequency synthesizer output should be an impulse at the desired frequency. In this section, the factors affecting the quality of the synthesized signal as well as the sources of phase noise within a phase-locked loop and a DDFS are examined.

### 12.4.1 Definition of Phase Noise

The most prevalent definition of phase noise is expressed in terms of  $L(f_m)$ , which is the normalized frequency domain representation of phase fluctuations. It is the ratio of the power spectral density (PSD) in one phase modulation sideband, referred to the carrier frequency on a spectral density basis, to the total signal power, at a frequency offset  $f$ . The units for this quantity are  $\text{Hz}^{-1}$ . This quantity,  $L(f_m)$ , is a two-sided spectral density [114]. It is also called single sideband phase noise [126]. Note that, single-sided signals do not correspond to single-sided spectral densities. They are totally distinct sets of terminology. Single-sided signals only refers to the nature of the signal; whereas, single-sided spectral densities refers to how the signal is represented.

Another important definition of phase noise is the single-sided power spectral density definition,  $S_\phi(f)$ . The dimensions of this quantity are  $\text{rad}^2/\text{Hz}$  [114]. Note that if the total phase noise energy is much less than  $1 \text{ rad}^2$ , the value of

the one-sided spectral density representation of the phase noise is half of that of the double-sided representation of phase noise.

### 12.4.2 Basic Mechanism of Phase Noise

Consider a sine wave corrupted by phase noise given as:

$$V(t) = V \sin[\omega_t t + n_2(t)] \quad (12.11)$$

where,  $n_2(t)$ , in general, is some random process. To simplify the analysis let  $n_2(t)$  be a sine function. Therefore, equation (12.11) becomes:

$$V(t) = V \sin[\omega_t t + \Delta f/f_m \sin w_m t] \quad (12.12)$$

where,  $\Delta f$  is the peak frequency deviation and  $\theta_p = \Delta f/f_m$  is the peak phase deviation, which is also known as the modulation index. Expanding equation (12.12) yields:

$$V(t) = V [\cos(\omega_t t) \cos(\theta_p \sin(\omega_t t)) - \sin(\omega_t t) \sin(\theta_p \sin(\omega_t t))] \quad (12.13)$$

Using the Bessel function  $V(t)$  can be expressed as:

$$\begin{aligned} V(t) &= V \cos(w_m t) \left[ J_0(\theta_p) + 2 \sum_{k=2i} J_k(\theta_p) \cos(k\omega_t t) \right] \\ &\quad - \left[ 2 \sum_{k=2i} J_k(\theta_p) \sin(k\omega_t t) \right] \end{aligned} \quad (12.14)$$

If it is assumed that  $\theta_p \ll 1$  (i.e. small phase noise), then:

$$\cos(\theta_p \sin(\omega_t t)) \approx 1 \quad (12.15)$$

and

$$\sin(\theta_p \sin(\omega_t t)) \approx \theta_p \sin(\omega_t t) \quad (12.16)$$

Therefore, equation (12.14) can be simplified to:

$$\begin{aligned} V(t) &= V [\cos(\omega_t t) - \sin(\omega_t t)(\theta_p \sin(\omega_t t))] \\ &= V \{\cos(\omega_t t) - (\theta_p/2)[\cos(\omega_t - \omega_m)t \\ &\quad - \cos(\omega_t + \omega_m)t]\} \end{aligned} \quad (12.17)$$

This means that as long as the frequency deviation (or alternatively the phase noise) of the frequency synthesizer is small, the output spectrum yields a single pulse at the desired frequency and two sidebands at frequencies  $\omega_t \pm \omega_m$  and their amplitude is equal to  $V * \theta_p/2$ . The single sideband phase noise is given as:

$$L(f_m) = (V_{\text{noise}}/V_{\text{signal}})^2 = \theta_p^2/4 = \theta_{\text{rms}}^2/2 \quad (12.18)$$

Since it was assumed that  $\theta_p \ll 1$ , the double sideband phase noise is given as:

$$S_\phi(f) = 2L(f_m) = \theta_p^2/2 = \theta_{\text{rms}}^2 \quad (12.19)$$

It is important to remember that when measuring these quantities,  $\theta_p$  and  $\theta_{\text{rms}}$  are both measured over a 1-Hz bandwidth.

### 12.4.3 Phase Noise in Fractional-N PLL Synthesizers

There are two types of phase noise in PLL synthesizers. The first is long term phase noise, usually caused by the external frequency reference source, and the other is short term phase noise, usually contributed by the PLL itself [127]. Most phase noise present in the PLL is caused by the VCO and the frequency divider. Figure 12.14 demonstrates the effect of the PLL loop on the phase noise of the VCO. It seems that the closed loop PLL reduces the phase noise of the VCO up to its bandwidth. To understand how this works, consider the simple PLL model shown in Figure 11.1 of chapter 11 with the VCO noise taken into account. If the input is taken to be from the VCO noise source and the output from the VCO output, the transfer function becomes:

$$H_2(jw) = \frac{1}{1 + K_d K_o H_{LF}(jw)/(jNw)} \quad (12.20)$$

Using  $H_{LF}(jw)$  as given by equation 11.5 of chapter 11,  $H_2(jw)$  becomes:

$$H_2(jw) = \frac{N}{K_o K_d K_f} \frac{\tau_f(jw)^2 + jw}{\frac{N\tau_f}{K_o K_d K_f}(jw)^2 + \frac{N}{K_o K_d K_f} jw + 1} \quad (12.21)$$

This is clearly a high pass filter. The denominator of  $H_2(jw)$  is the same as  $H_{PLL}(jw)$ , which is given in equation 11.9 of chapter 11. This means that the bandwidth of the high pass filter is equal to the bandwidth of the PLL, which is consistent with Figure 12.14. The larger the PLL bandwidth, the more VCO generated phase noise is suppressed. This makes sense since the larger the PLL's bandwidth, the faster it can correct for any phase deviations that the

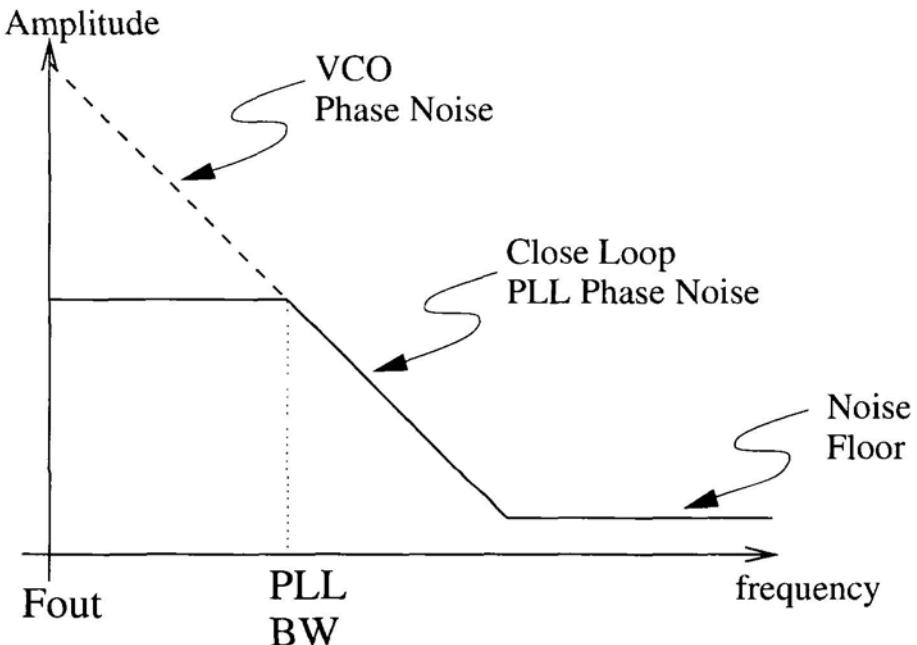


Figure 12.14. Phase noise in a closed-loop PLL system.

VCO produces and hence the magnitude of the phase deviations are limited. This type of phase noise is referred to as short-term phase noise.

The crystal oscillator driving the PLL generates another, less common, source of phase noise. Noise generated by the crystal oscillator is shaped by the transfer function  $H(s)$ , which is generally a low pass filter. Therefore, there is a trade-off in choice of the PLL bandwidth. Larger bandwidth leads to more phase noise suppression generated by the VCO, but allows more phase noise generated by the crystal oscillator. For integrated PLLs, most of the phase noise is generated by the VCO and hence the bandwidths tend to be larger.

There is another source of phase noise, which is indigenous to fractional-N PLLs. This results from the non-ideal beat note cancellation of the fractional ratio at the input to the phase detector. Consider for example, that a fractional division ratio of 4.8 is required. To achieve this, the division ratio is held at 5 for four cycles, then 4 for one cycle. Over a large number of cycles, this would yield a fractional division ratio of 4.8. Practically, however, this ideal fractional division ratio is not reached quickly. There is always a residual division ratio error, which would lead to random phase cancellation errors at the phase detector. Let  $a_n$  represent this random phase cancellation errors. The noise terms may now be represented as [1]:

$$n(t) = \sum_{n=-\infty}^{\infty} a_n \text{rect}(t - nT) \quad (12.22)$$

The power spectral density then becomes:

$$\begin{aligned} S_n(f) &= \lim_{M \rightarrow \infty} \frac{1}{2MT} E \left\{ |F[n(t)]|^2 \right\} \\ &= \frac{\sigma^2}{F_{ref}} \left[ \frac{\sin(\pi f/F_{ref})}{(\pi f/F_{ref})} \right]^2 \text{ rad}^2/\text{Hz} \end{aligned} \quad (12.23)$$

where,  $\sigma^2$  is the variance of the random cancellation errors. Note that, this result shows that as the reference frequency increases, this contribution of phase noise decreases. This is intuitive, since higher reference frequency leads to more random error sequence samples averaging of a fixed time interval.

#### 12.4.4 Phase Noise in DDFS Systems

Phase noise in DDFS systems is simpler in comparison to PLLs. Since the DDFS is entirely digital system, the DDFS itself does not add to the overall phase noise in the system. The main source of the phase noise in the DDFS system is contributed by the clock signal itself. Any errors in the clock signal result are translated directly to phase noise at the output. This is because both the digital circuitry and the digital-to-analog converter are controlled by the clock signal. Any phase errors in the clock signal would result in incorrect timing in generating a new signal from the digital-to-analog converter, resulting in phase noise.

As stated previously, phase noise is the normalized frequency domain representation of phase fluctuations. This normalization is with respect to the instantaneous period of oscillation. Since the clock period is always higher than the output period of the DDFS, the phase noise output of the DDFS should be lower than the clock period. The difference in phase noise is equal to the ratio of the output frequency to the clock frequency.

#### 12.5 SUMMARY

In this chapter, the basic frequency synthesis techniques were introduced. Important frequency synthesizer parameters have been given. Various system level design factors that affect the frequency synthesizer parameters such as modulation technique, transmission power, and operating frequency range have been described in relation to the frequency synthesis problem. The most important of these parameters is the phase noise of the frequency synthesizer. Phase noise has been shown to be different from amplitude noise in that phase noise refers to signal timing errors as opposed to amplitude errors. Phase noise

is also affected by system level design factors such as the choice of modulation scheme and output power.

The two basic frequency synthesizer techniques used today are the phase-lock loop (PLL) and direct digital frequency synthesizer (DDFS) techniques. PLL synthesis is an indirect type of frequency synthesis in which a low frequency reference signal is used to produce a high quality high frequency signal. This is through the use of a feedback control mechanism with a frequency divider in the feedback path. The division ratio of this frequency divider determines the multiplication ratio.

An improvement over the basic PLL technique is the fractional-N PLL synthesizer. It uses two feedback division ratios instead of one. These two division ratios would be dynamically altered from one to the other to produce an average division ratio, which lies between the two division ratios. This helps to lower the values of the feedback division ratio as well as increase the PLL's reference frequency. This has the effect of increasing the PLL's loop bandwidth. This has three important implications. Firstly, since the loop bandwidth of the PLL is increased, the response time is decreased and, hence, the settling time of the PLL becomes faster. Secondly, since the PLL "cleans" the VCO's phase noise over a range extending over its closed loop bandwidth, more of the VCO's phase noise is suppressed. Thirdly, since the feedback division ratio of the PLL is decreased, the  $20 \log(N)$  degradation in noise floor is now less and hence the noise floor of the PLL is improved.

The main disadvantage of the fractional-N PLL technique is the introduction of spurs. The spurs are created due to the instantaneous switching of feedback division ratios. One proposed means of removing these spurs is by the introduction of the digital  $\Sigma - \Delta$  modulator to randomize the division ratio transitions of the feedback frequency divider. Such "randomization" transfers the noise (or spur) to high frequencies, which can then be filtered by the closed loop bandwidth of the PLL.

The second method of frequency synthesis is direct digital frequency synthesis (DDFS). This technique is a digital-based solution to frequency synthesis. Its main advantages are that it is a digital implementation (i.e. easily repeatable and amenable to VLSI integration), it exhibits instantaneous lock time; it can perform linear phase shifting (important in PSK modulation), and it adds no phase noise to the reference clock signal. Its main disadvantages include limited operating frequency, higher power consumption (relative to PLLs), and higher level of spurs (relative to PLLs).

The problem of spur reduction has been the subject to intensive research. The sources of spurs have been identified to be caused by finite word length effects in the phase accumulator, ROM lookup table, and also from DAC non-linearities. The spurs caused by DAC non-linearities are the dominant cause of spurs in DDFS.

One important frequency synthesis parameter is phase noise. The most prevalent definition of phase noise is expressed in terms of  $L(f_m)$ , which is the normalized frequency domain representation of phase fluctuations. It is the ratio of the power spectral density (PSD) in one phase modulation sideband, referred to the carrier frequency on a spectral density basis, to the total signal power, at a frequency offset  $f$ . The units for this quantity are  $\text{Hz}^{-1}$ .

Phase noise arises from timing errors in the frequency synthesizer. In the frequency domain this results in smearing of the synthesized signal to a wider range of frequencies than desired. The fractional-N PLL has an additional source of phase noise as a result of the non-ideal beat note cancellation of the fractional ratio at the input to the phase detector. This source of phase noise decreases with increasing the reference frequency since the higher reference frequency leads to more random error sequence samples averaging over a fixed time interval.

Finally, the phase noise of DDFS has been analyzed. As stated earlier, the DDFS does not add any phase noise to the output spectrum. On the contrary, since the output frequency of the DDFS is less than its reference clock frequency, the phase noise is reduced.

## References

- [1] J. Crawford. *Frequency Synthesizer Design Handbook*. Artech House, Boston, 1994.
- [2] A. Fahim and M. Elmasry. A low-power frequency synthesizer design methodology for wireless applications. In *IEEE Int'l Symposium on Circuits and Systems*, 1998.
- [3] M. Sakamoto. Analog cellular radio in japan. In D. M. Balston and R. C. V. Macario, editors, *Cellular Radio Systems*, chapter 5, pages 135--149. Artech House, Boston, 1993.
- [4] ARIB. Radio system overview of 3g mobile system, September 1998.
- [5] Tsuneo Tsukkahara, Masayuki Ishikawa, and Masahiro Muraguchi. A 2V 2GHz Si-bipolar direct conversion quadrature modulator. In *IEEE International Solid-State Circuits Conference*, pages 40--41, San Francisco, California, February 1994.
- [6] Gordon White. *Mobile Radio Technology*. Newnes, Oxford, England, 1994.
- [7] Jan van Duuren, Peter Kastelein, and Frits C. Schoute. *Fixed and Mobile Telecommunications: Networks Systems and Services*. Addison-Wesley, Harlow, England, second edition, 1996.
- [8] I. Korn. *Digital Communications*. Van Nostrand Reinhold Company, New York, 1985.
- [9] Michael L. Hinig and Melbourbe Barton. Baseband signalling and pulse shaping. In Jerry D. Gibson, editor, *The Mobile Communications Handbook*, chapter 4, pages 35--55. CRC Press, 1996.
- [10] Kamilo Feher. *Digital Communications: Satellite/Earth Station Engineering*. Prentice-Kall, Englewood Cliffs, NJ, 1981.
- [11] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379--423,1949.
- [12] S. Lin and D.J. Costello. *Error Control Coding: Fundamentals and Applications*. Prentice-Kall, Englewood Cliffs, New Jersey, 1983.
- [13] G.C. Clark and J.B. Cain. *Error-Correction Coding for Digital Communications*. Plenum, New York, 1981.

- [14] A. Bruce Carlson. *Communications Systems: An Introduction to Signal and Noise in Electrical Communication*. McGraw-Hill, 1986.
- [15] William C. Y. Lee. *Mobile Cellular Telecommunications: Analog and Digital Systems*. McGraw-Hill, 1995.
- [16] Emad N. Farag, Mohamed I. Elmasry, Mohamed N. Saleh, and Nabil M. Elnady. A two-level hierarchical mobile network: Structure and network control. *International Journal of Reliability, Quality and Safety Engineering*, 3:325--351, December 1996.
- [17] R. Steele. The cellular environment of lightweight handheld portables. *IEEE Communications Magazine*, 27:20--29, July 1989.
- [18] Theodore S. Rappaport. *Wireless Communications: Principles and Practice*. Prentice Hall, Upper Saddle River, New Jersey, 1996.
- [19] Seiichi Sampei. *Applications of Digital Wireless Technologies to Global Wireless Communications*. Prentice Hall, Upper Saddle River, New Jersey, 1997.
- [20] Dale R. Carson and Paul Lorrain. *Introduction to Electromagnetic Fields and Waves*. W. H. Freeman and Company, San Francisco, 1967.
- [21] D. C. Cox, R. R. Murray, and A. W. Norris. 800-MHz attenuation measured in and around suburban houses. *AT&T Bell Laboratories Technical Journal*, 63:921--954, Jul--Aug. 1984.
- [22] Bernard Sklar. Rayleigh fading channels in mobile digital communication systems part I: Characterization. *IEEE Communications Magazine*, 35:136--146, September 1997.
- [23] Richard C. Bernhardt. Macroscopic diversity in frequency reuse radio systems. *IEEE Journal on Selected Areas in Communications*, SAC-5:862--870, June 1987.
- [24] Raymond Steele. *Mobile Radio Communications*. Pentech Press, London, 1992.
- [25] William C. Jakes. *Microwave Mobile Communications*. IEEE Press, Piscataway, New Jersey, 1993.
- [26] D. Greenwood and L. Hanzo. Characterisation of mobile radio channels. In Raymond Steele, editor, *Mobile Radio Communications*, chapter 2, pages 92--185. Pentech Press, London, 1992.
- [27] Leon W. Couch. *Digital and Analog Modulation Systems*. Prentice Hall, Upper Saddle River, New Jersey, 1997.

- [28] Sandeep Chennakeshu and Gray J. Saulnier. Differential detection of  $\pi/4$ -shifted-DQPSK for digital cellular radio. *IEEE Transactions on Vehicular Technology*, 42:46--57, February 1993.
- [29] F.G. Jenks, P.O. Morgan, and Christopher S. Warren. Use of four-level phase modulation for digital mobile radio. *IEEE Transactions on Electromagnetic Compatibility*, EMC-14:113--128, November 1972.
- [30] Yanpen Guo and Kamilo Feher. Modem/radio IC architectures for ISM band wireless applications. *IEEE Transactions on Consumer Electronics*, 39:100-106, May 1993.
- [31] Steven H. Goode, Henry L. Kazecki, and Donald W. Dennis. A comparison of limiter-discriminator, delay and coherent detection for  $\pi/4$  QPSK. In *IEEE Vehicular Technology Conference*, pages 687--694, Orlando, Florida, May 1990.
- [32] Yoshihiko Akaiwa and Yoshinori Nagata. Highly efficient digital mobile communications with a linear modulation method. *IEEE Journal on Selected Areas in Communications*, SAC-5:890--895, June 1987.
- [33] P. T. Brady. A statistical analysis of on-off patterns in 16 conversations. *Bell System Technical Journal*, 47:73--91, 1968.
- [34] Andrew J. Viterbi. *CDMA: Principles of Spread Spectrum Communication*. Addison-Wesley Publishing Company, 1995.
- [35] Vijay K. Garg, Kenneth F. Smolik, and Joseph E. Wilkes. *Applications of CDMA in Wireless/Personal Communications*. Prentice Hall, Upper Saddle River, New Jersey, 1997.
- [36] Jan Crols and Michiel Steyaert. *CMOS Wireless Transceiver Design*. Kluwer Academic Publishers, Boston, 1997.
- [37] Joe Mitola. The software radio architecture. *IEEE Communications Magazine*, pages 26--38, May 1995.
- [38] Raymond J. Lackey and Donald W. Upmal. Speakeasy: The military software radio. *IEEE Communications Magazine*, pages 56--61, May 1995.
- [39] Joseph Kennedy and Mark C. Sullivan. Direction finding and “smart antennas” using software radio architectures. *IEEE Communications Magazine*, pages 62--68, May 1995.
- [40] Jeffery A. Wepman. Analog-to-digital converters and their applications in radio receivers. *IEEE Communications Magazine*, pages 39--45, May 1995.

- [41] Rupert Baines. The DSP bottleneck. *IEEE Communications Magazine*, pages 46--54, March 1995.
- [42] Alan J. Coulson. A generalization of nonuniform bandpass sampling. *IEEE Transactions on Signal Processing*, 43:694--704, March 1995.
- [43] Heinrich Meyr and Ravi Subramanian. Advanced digital receiver principles and technologies for PCS. *IEEE Communications Magazine*, pages 68--78, January 1995.
- [44] Allen Gersho and Robert M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1992.
- [45] B.S. Song, S.H. Lee, and M.F. Tompsett. A 10-bit 15 MHz COMS recycling two-step A/D converter. *IEEE Journal of Solid State Circuits*, 25:1238--1338, December 1990.
- [46] A. G. Dingwall and V. Zazzu. An 8-MHz CMOS subranging 8-bit A/D converter. *IEEE Journal of Solid State Circuits*, 20:1138--1143, December 1995.
- [47] R. Schreier and M. Snelgrove. Bandpass sigma-delta modulation. *Electronics Letter*, 25:1560--1561, November 1989.
- [48] Stephen A. Jantzi, W. Martin Snelgrove, and Paul F. Ferguson. A fourth-order bandpass sigma-delta modulator. *IEEE Journal of Solid State Circuits*, 28:282--291, March 1993.
- [49] Frank W. Singorand W. Martin Snelgrove. Switched-capacitor bandpass delta-sigma A/D modulation at 10.7 MHz. *IEEE Journal of Solid State Circuits*, 30:184--192, March 1995.
- [50] James C. Candy. Decimation for sigma delta modulation. *IEEE Transactions on Communications*, COM-34:72--76, January 1986.
- [51] Brian P. Brandt and Bruce A. Wooley. A low-power area efficient digital filter for decimation and interpolation. *IEEE Journal of Solid State Circuits*, 29:679--687, June 1994.
- [52] Asad A. Abidi. Low-power radio-frequency ic's for portable communications. *Proceedings of the IEEE*, 83:544--569, April 1995.
- [53] Anantha P. Chandrakasan, Samuel Sheng, and Robert W. Brodersen. Low-power CMOS digital design. *IEEE Journal of Solid State Circuits*, SC-27:473--484, April 1992.

- [54] John L. Hennessy and David A. Patterson. *Computer Architecture A Qualitative Approach*. Morgan Kaufmann Publishers, Inc., San Mateo, California, 1990.
- [55] Behzad Razavi. Recent advances in RF integrated circuits. *IEEE Communications Magazine*, 35:36--43, December 1997.
- [56] Jan M. Rabaey and Massoud Pedram. *Low power design methodologies*. Kluwer Academic Publishers, Boston, 1996.
- [57] Abdellatif Bellaour and Mohamed I. Elmasry. *Low-Power Digital VLSI Design: Circuits and Systems*. Kluwer Academic Publishers, Boston, 1995.
- [58] Peter Michel, Ulrich Lauther, and Peter Duzy. *The Synthesis Approach to Digital System Design*. Kluwer Academic Publishers, Boston, 1992.
- [59] Deo Singh, Jan M. Rabaey, Massoud Pedram, Franck Catthoor, Suresh Rajgopal, Naresh Sehgal, and Thomas J. Mozdzen. Power conscious CAD tools and methodologies: A perspective. *Proceedings of the IEEE*, 83:570--594, April 1995.
- [60] Brian Nadel. The green machine. *PC Magazine*, 12:110--145, May 25 1993.
- [61] Teresa H. Meng, Benjamin M. Gorgon, Ely K. Tsern, and Andy C. Hung. Portable video-on-demand in wireless communication. *Proceedings of the IEEE*, 83:659--679, April 1995.
- [62] Samuel Sheng, Anantha Chandrakasan, and Robert W. Brodersen. A portable multimedia terminal. *IEEE Communications Magazine*, pages 64--75, December 1992.
- [63] Peter Pirsch, Nicolas Demassieux, and Winfried Gehrke. VLSI architectures for video compression --- A survey. *Proceedings of the IEEE*, 83:220--246, February 1995.
- [64] Anantha P. Chandrakasan and Robert W. Brodersen. Minimizing power consumption in digital CMOS circuits. *Proceedings of the IEEE*, 83:498--523, April 1995.
- [65] H. J. M. Veendrick. Short-circuit dissipation of static CMOS circuitry and its impact on the design of buffer circuits. *IEEE Journal of Solid State Circuits*, SC-19:468--473, August 1984.
- [66] Edward S. Yang. *Micro-Electronic Devices*. McGraw Hill, New York, 1988.

- [67] Jean-Pierre Collinge. *Silicon-on-Insulator Technology: Materials to VLSI*. Kluwer Academic Publishers, Boston, 1991.
- [68] S. M. Kang. Accurate simulation of power dissipation in VLSI circuits. *IEEE Journal of Solid State Circuits*, SC-21:889–891, October 1986.
- [69] Farid N. Najm. A survey of power estimation techniques in VLSI circuits. *IEEE Transactions on VLSI Systems*, 2:446–455, December 1994.
- [70] Mehmet A. Cirit. Estimating dynamic power consumption of CMOS circuits. In *IEEE International Conference on Computer-Aided Design*, pages 534–537, Santa Clara, California, November 1987.
- [71] A. Shen, Abhijit Ghosh, Srinivas Devadas, and Kurt Keutzer. On average power dissipation and random pattern testability of CMOS combinational logic networks. In *IEEE/ACM International Conference on Computer-Aided Design*, pages 402–407, Santa Clara, California, November 1992.
- [72] Farid N. Najm. Transition density: A new measure of activity in digital circuits. *IEEE Transactions on Computer-Aided Design*, 12:310–323, February 1993.
- [73] Abhijit Ghosh, Srinivas Devadas, Kurt Keutzer, and Jacob White. Estimation of average switching activity in combinational and sequential circuits. In *IEEE/ACM International Conference on Computer-Aided Design*, pages 253–259, Santa Clara, California, November 1992.
- [74] Anantha P. Chandrakasan, Miodrag Potkonjak, Renu Mehra, Jan Rabaey, and Robert W. Brodersen. Optimizing power using transformations. *IEEE Transactions on Computer-Aided Design*, 14:12–31, January 1995.
- [75] P. E. Landman and J. M. Rabaey. Power estimation for high level synthesis. In *Proceedings of the European Conference on Design Automation*, pages 304–308, Paris, France, February 1993.
- [76] Eric A. Vittoz. Low-power design: Ways to approach the limits. In *IEEE International Solid-State Circuits Conference*, pages 14–18, San Francisco, California, February 1994.
- [77] Donald C. Cox. Universal digital portable radio communications. *Proceedings of the IEEE*, 75:436–477, April 1987.

- [78] Akira Matsuza. Low-voltage and low-power circuit design for mixed analog/digital systems in portable equipments. *IEEE Journal of Solid State Circuits*, 29:470--480, April 1994.
- [79] Anantha P. Chandrakasan, Andrew Burstein, and Robert W. Brodersen. A low-power chipset for a portable multimedia I/O terminal. *IEEE Journal of Solid State Circuits*, 29:1415--1428, December 1994.
- [80] Thomas Barber, Phil Carvey, and Anantha Chandrakasan. Designing for wireless LAN communications. *Circuits & Devices Magazine*, pages 29--33, July 1996.
- [81] Inyup Kang and Alan N. Willson. A low-power state-sequantional Viterbi decoder for CDMA digital cellular applications. In *IEEE International Circuits and Systems Conference*, volume 4, pages 272--276, Atlanta, Georgia, May 1996.
- [82] R. Cypher and C. B. Shung. Generalized traceback techniques for survivor memory management in Viterbi decoder. In *GLOBECOM*, volume 2, pages 1318--1322, Houston, Texas, December 1993.
- [83] Emmanuel Boutillon and Nicolas Demassieux. High speed low power architecture for memory management in a Viterbi decoder. In *IEEE International Circuits and Systems Conference*, volume 4, pages 284--287, Atlanta, Georgia, May 1996.
- [84] Mark D. Hahm, Eby G. Friedman, and Edward L. Titlebaum. Analog vs. digital: A comparison of circuit implementations for low-power matched filters. In *IEEE International Circuits and Systems Conference*, volume 4, pages 280--283, Atlanta, Georgia, May 1996.
- [85] Anantha P. Chandrakasan, Miodrag Potkonjak, Jan Rabaey, and Robert W. Brodersen. HYPER-LP: A system for power minimization using architectural transformations. In *IEEE International Conference on Computer-Aided Design*, pages 300--303, Santa Clara, California, November 1992.
- [86] Wai Lee et al. A IV DSP for wireless communications. In *IEEE International Solid-State Circuits Conference*, pages 92--93, San Francisco, California, February 1997.
- [87] Tadahiro Kuroda et al. A 0.9V 150MHz 10mW 4mm<sup>2</sup> 2-D discrete cosine transform core processor with variable-threshold-voltage scheme. In *IEEE International Solid-State Circuits Conference*, pages 166--167, San Francisco, California, February 1996.

- [88] Johannes M. C. Stork. Technology leverage for ultra-low power information systems. *Proceedings of the IEEE*, 83:607–618, April 1995.
- [89] Masayuki Ino et al.  $0.25\mu\text{m}$  CMOS/SIMOX gate array LSI. In *IEEE International Solid-State Circuits Conference*, pages 86–87, San Francisco, California, February 1996.
- [90] Masakazu Kakumu and Masaaki Kinugawa. Power-supply voltage impact on circuit performance for half and lower submicrometer CMOS LSI. *IEEE Transactions on Electron Devices*, 37:1902–1908, August 1990.
- [91] S. Mutoh et al. 1 V high-speed digital circuit technology with  $0.5\text{-}\mu\text{m}$  multi-threshold CMOS. In *IEEE International ASIC Conference and Exhibit*, pages 186–189, Rochester, New York, September 1993.
- [92] Kasuo Yano et al. A 3.8-ns CMOS  $16 \times 16\text{-b}$  multiplier using complementary pass-transistor logic. *IEEE Journal of Solid State Circuits*, 25:388–395, April 1990.
- [93] Anthony J. Stratakos, Seth R. Sanders, and Robert W. Bordersen. A low-voltage CMOS DC-DC converter for a portable battery operated system. In *IEEE Power Electronics Specialists Conference*, pages 619–626, 1994.
- [94] M. Alidina, J. Monterio, S. Devadas, A. Ghosh, and M. Papaefthmiou. Precomputation-based sequential logic optimization for low power. In *International Workshop in Low Power Design*, 1994.
- [95] Takakuni Douseki et al. A 0.5V SIMOX-MTCMOS circuit with 200ps logic gate. In *IEEE International Solid-State Circuits Conference*, pages 84–85, San Francisco, California, February 1997.
- [96] Srinivasa R. Vemuru and Arthur R. Thorbjornsen. Variable-taper CMOS buffer. *IEEE Journal of Solid State Circuits*, 26:1265–1269, September 1991.
- [97] Thomas K. Callaway and Earl E. Swartzlander. Optimizing arithmetic elements for signal processing. In *VLSI Signal Processing Workshop*, pages 91–100, 1992.
- [98] Emad N. Farag and Mohamed I. Elmasry. Using carry save adders to reduce power dissipation. In *Eighth International Conference on Microelectronics*, pages 173–176, Cairo, Egypt, December 1996.
- [99] Ching-Long Su, Chi-Ying Tsui, and Alvin M. Despain. Low power architecture design and compilation techniques for high-performance

- processors. In *Compcon*, pages 489--498, San Francisco, California, March 1994.
- [100] D. Wong, B. H. Juang, and A. H. Gray. An 800 bit/s vector quantization LPC vocoder. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-30:770--779, October 1982.
  - [101] David A. Hodges and Horace G. Jackson. *Analysis and Design of Digital Integrated Circuits*. McGraw Hill, New York, 1988.
  - [102] Paul R. Gary and Robert G. Meyer. *Analysis and Design of Analog Integrated Circuits*. John Wiley & Sons, 1993.
  - [103] A. Asensio and F. Perez. Design and optimization of high-power plused class-C microwave amplifiers. In *IEEE International Symposium on Circuits and Systems*, pages 2371--2374, New York, 1991.
  - [104] R. Best. *Phase-locked Loops: Theory, Design, and Applications*. McGraw-Hill, New York, 1993.
  - [105] M. Manassewitsch. *Frequency Synthesizers Theory and Design*. John Wiley & Sons, New York, NY, 1980.
  - [106] Lawrence E. Larson. *RF and Microwave Circuit Design for Wirless Communications*. Artech House, Boston, 1996.
  - [107] Jacob Millman and Christos C. Halkias. *Electronic Devices and Circuits*. McGraw-Hill, New York, 1967.
  - [108] Adel S. Sedra and Kenneth C. Smith. *Microelectronic Circuits*. Saunders College Publishing, 1991.
  - [109] David J. Comer. *Modern Electronic Circuit Design*. Addison Wesley, Reading, MA, 1976.
  - [110] Ulrich L. Rohde and T.T. Nelson Bucher. *Communications Receivers: Principles & Design*. McGraw Hill, New York.
  - [111] R.E. Senz and R.A. Bartkowiak. *Feedback amplifiers and oscillators*. 1968.
  - [112] Benjamin Parzen. *Design of Crystal and Other Harmonic Oscillators*. John Wiley & Sons, 1983.
  - [113] V. Kroupa. *Frequency Synthesis: Theory, Design & Applications*. Charles Griffin & Company Ltd., London, Great Britain, 1973.
  - [114] B. Goldberg. *Digital Techniques in Frequency Synthesis*. McGraw-Hill, New York, NY, 1996.

- [115] W. Djen. Fractional-N PLL provides fast, low-noise synthesis. *Microwaves & RF*, pages 95--102, May 1994.
- [116] M. Mano. *Digital Design*. Prentice-Hall, Englewood Cliffs, NJ, 1991.
- [117] T. Riley et al. Delta-sigma modulation in fractional-N frequency synthesis. *IEEE Journal of Solid State Circuits*, pages 533--539, May 1993.
- [118] B. Miller. Technique enhances the performance of PLL synthesizers. *Microwaves & RF*, pages 59--65, January 1993.
- [119] B. Leung. The oversampled technique for A/D conversion: A tutorial overview. *Analog Integrated Circuit and Signal Processing I*, pages 65--74, 1991.
- [120] J. Vankka. Spur reduction techniques in sine output direct digital synthesis. In *IEEE International Frequency Control Symposium*, pages 951--959, 1996.
- [121] J. Vankka. Methods of mapping from phase to sine amplitude in direct digital synthesis. In *IEEE International Frequency Control Symposium*, pages 942--950, 1996.
- [122] L. Kushner. The composite DDS - a new direct digital synthesizer architecture. In *IEEE International Frequency Control Symposium*, pages 255--260, 1993.
- [123] V. Kroupa. Discrete spurious signals and background noise in direct digital frequency synthesizers. In *IEEE International Frequency Control Symposium*, pages 242--250, 1993.
- [124] V. Reinhardt. Spur reduction techniques in direct digital synthesizers. In *IEEE International Frequency Control Symposium*, pages 230--241, 1993.
- [125] L. Kushner. A spurious reduction technique for high-speed direct digital synthesizers. In *IEEE International Frequency Control Symposium*, pages 920--927, 1996.
- [126] U. Rohde. *Digital PLL Frequency Synthesizers: Theory and Design*. Prentice Hall, Englewood Cliffs, N.J., 1983.
- [127] D. Wolaver. *Phase-locked Loop Circuit Design*. Prentice-Hall, Englewood Cliffs, N.J., 1991.

# Index

- $\pi/4$  - DQPSK, 103
  - space diagram, 104
- Advanced Mobile Phone Service, 2, 9
  - amplitude shift keying, 88
    - bit error rate, 89
    - power spectral density, 88
- AMPS, 2, 9
  - analog-to-digital converters, 155
    - performance metrics, 158
  - application specific integrated circuits, 204
  - ASK, 88
  - automatic gain control amplifiers, 238
  - band-pass sampling, 165
  - band-pass Sigma-Delta modulator, 192
  - Barkhausen criterion, 273
  - baseband subsystem, 196, 199
    - battery capacity, 211
  - BiCMOS, 209
  - binary phase shift keying, 91
    - bit error rate, 92
    - power spectral density, 91
  - bipolar junction transistor, 228
    - high-frequency pi-model, 229
    - unity current gain frequency, 231
  - block codes, 38
  - block interleaving, 42
  - BPSK, 91
    - brick wall filter, 28
      - impulse response, 28
  - CDMA, 15
  - cell clustering, 57
  - cell shape, 56
  - cell splitting, 63
  - channel capacity, 37
  - channel coding, 37
  - channel encoder, 24
  - channel impairments, 23, 66
  - characteristic equation, 124
  - Clapp oscillator, 281
  - CMOS, 209, 213
    - co-channel interference, 57
    - Colpitts oscillator, 275
    - convolution codes, 39
    - crystal oscillators, 281
  - DBPSK, 93
  - decimation filter, 192
  - demodulation, 26
  - differential amplifier, 238
  - differential BPSK, 93
    - bit error rate, 95
  - differential non-linearity, 158
  - differential QPSK, 99
    - bit error rate, 103
  - digital receiver, 25
  - digital signal processors, 199
  - digital transmitter, 24
  - digital-to-analog converters, 156
  - direct conversion receiver, 206
  - direct digital frequency synthesizers, 297
  - direct sequence spread spectrum, 119
  - Doppler shift, 72
  - DQPSK, 99
  - dynamic range, 198
  - effective isotropic radiated power, 68
  - ETACS, 10
  - Extended Total Access Communication System, 10
  - fast fading, 77
  - feedback amplifier, 273
  - Fibonacci feedback shift register, 122
  - field effect transistor, 228
    - high-frequency pi-model, 232
    - unity current gain frequency, 232

- field programmable gate arrays, 201
- first-order Sigma-Delta modulator, 189, 295
- Hash analog-to-digital converters, 171
- flat fading, 77
- flat-top sampling, 164
- flicker noise, 234
- folding analog-to-digital converters, 182
- forward error correction, 37
- fractional-N phase locked loop, 289
- free space propagation, 67
- frequency diversity, 85
- frequency dividers, 269
- frequency hopping spread spectrum, 119
- frequency selective fading, 75
- frequency synthesizers, 285
- Galois feedback shift register, 122
- Gaussian minimum shift keying, 111
- Global System for Mobile communications, 2, 11
- GMSK, 111
- GSM.2, 11, 205
- Hamming distance, 38
- handoff, 64
- hardware design languages, 201
- Hartley oscillator, 275
- homodyne receiver, 7, 150, 206
- ideal sampling, 159
- image frequency, 145
- Improved Mobile Phone System, 1
- IMT-2000, 3
- integral non-linearity, 158
- Inter-symbol Interference, 28
- interleaving, 41
- intermodulation distortion, 138
- International Mobile Telecommunications 2000, 3
- interpolative analog-to-digital converters, 173
- irreducible primitive polynomial, 124
- IS-136, 2, 12
- IS-95, 15, 205
- isotropic transmitter, 67
- jitter, 35
- Johnson noise, 233
- JTACS, 11
- large scale fading, 71
- leakage-current power dissipation, 214
- linear block codes, 38
- log-normal distribution, 71
- loop filter, 256
- low noise amplifier, 7, 209, 233
- low-power design, 211
- matched filter, 52
- maximal length sequence, 124
- mid-rise quantizer, 167
- mid-tread quantizer, 167
- minimum shift keying, 106
  - bit error rate, 110
  - power spectral density, 110
- mixers, 210
- Mobile Telephone System, 1
- modulation, 25, 87
- MSK, 106
- NAMPS, 11
- negative resistance oscillator, 278
- NMT, 10
- noise figure, 133, 234
- noise temperature, 134
- non-linearity, 138
- Nordic Mobile Telephone, 2, 10
- NTACS, 11
- Nyquist limit, 27
- Nyquist sampling condition, 162
- Nyquist's first theorem, 28
- offset QPSK, 99
- on-off keying, 88
- optimum receiver design, 45
- OQPSK, 99
- oscillator design, 273
- parallelism, 221
- Parseval's theorem, 51
- path diversity, 85
- path loss exponent, 72
- Personal Digital Cellular, 2
- phase detector, 256
  - analog, 261
  - digital, 262
- phase locked loops, 255
  - operation, 257
  - stability, 260
- phase noise, 287, 299
- phase-shift oscillator, 274
- pipelining, 220
- power amplifier, 210, 243
  - class A, 244
  - class B, 247
  - class C, 251
  - class F, 252
- power control, 130
- power dissipation estimation, 215
- primitive polynomial, 124
- processing gain, 117
- pseudorandom noise sequence, 121
- pulse sampling, 163
- QPSK, 96
- quadrature phase shift keying, 96

- bit error rate, 99
- power spectral density, 97
- quantization, 167
- raised cosine filter, 29
  - frequency response, 29
  - impulse response, 30
  - noise bandwidth, 30
  - roll-off factor, 30
- rake receiver, 85
- Rayleigh distribution, 79
- Rayleigh fading, 78
- reuse distance, 59
- Reverse-bias diode leakage current, 215
- RF subsystem, 196,204
- Ricean distribution, 83
- Ricean fading, 83
- sampling, 158
- Schwarz's inequality, 50
- second-order intercept point, 142
- second-order Sigma-Delta modulator, 190
- Shannon's limit, 36
- short-circuit-current power dissipation, 214
- shot noise, 234
- Sigma-Delta analog-to-digital converters, 188
- Sine decimator, 193
- single side-band mixer, 147
- slow fading, 77
- small scale fading, 75
- soft handoff, 65
- software radio, 151
- source decoder, 26
- sources of power dissipation, 213
- space diversity, 85
- spectrum allocations, 5
- spread spectrum, 115
- spread spectrum synchronization, 126
- square root raised cosine filter, 32
  - frequency response, 32
  - impulse response, 33
  - noise bandwidth, 33
- subranging analog-to-digital converters, 180
- Subthreshold current, 215
- successive approximation analog-to-digital converters, 186
- superheterodyne receiver, 7, 144, 206
- switching activity factor, 214
- switching power dissipation, 213
- systematic block code, 39
- TAGS, 2, 9, 11
- TDMA, 12
- thermal noise, 233
- Third generation mobile communication systems, 3
- third-order intercept point, 143
- time diversity, 85
- time hopping spread spectrum, 120
- Total Access Communication System, 2, 9
- tracking, 128
- transceiver design constraints, 197
- two step analog-to-digital converters, 177
- two-ray propagation model, 69
- UMTS, 3
- vector quantization, 223
- Viterbi algorithm, 41
- VLSI technology, 195
- voltage scaling, 198, 219
- voltage-controlled oscillator, 256, 284
- wireless transceiver design metrics, 26
- wireless transceivers, 6
- world administrative radio conference, 3