SATELLITE COMMUNICATIONS PAYLOAD AND SYSTEM

SATELLITE COMMUNICATIONS PAYLOAD AND SYSTEM

TERESA M. BRAUN

WILEY
A JOHN WILEY & SONS, INC., PUBLICATION

IEEE
IEEE PRESS

Cover image: KA-SAT drawing courtesy of Eutelsat, color by T. M. Braun

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data:

Braun, T. M., 1949–
Satellite Communications Payload and System/ T.M. Braun.
pages cm
ISBN 978-0-470-54084-8 (hardback)
1. Artificial satellites in telecommunication. 2. Artificial satellites—Electronic equipment.
3. Artificial satellites—Radio antennas. 4. High altitude platform systems (Telecommunication)
5. Digital communications. I. Title.

TK5104.B765 2012
621.3841′56—dc23

2012000091

Printed in the United States of America

10 9 8 7 6 5 4 3 2 1

CONTENTS

PART II PAYLOAD IN END-TO-END COMMUNICATIONS SYSTEM

PREFACE

This book is about the payload of communications satellites. Several books have been written about the satellite bus or platform, particular satellites, general satellite communications, and applications of satellite communications, but the payload has been covered only briefly in these books. Detailed books on how to design the various payload units, for example, antennas and filters, exist, but no book focuses on unit performance at a level appropriate for payload systems engineering.

This book is a unique combination of practical payload systems engineering and applied communications theory. As the payload is desired to have higher and higher capability, it becomes larger and more complex. More complicated analyses are required. The payload systems engineer is called upon to deal with things he never had to before or deal with the old things in a more exact way, to squeeze out higher performance. The engineer about to model the end-to-end communications system needs to fully understand the payload subtleties, new or old. The writer of the payload part of the proposal needs to realize that the formulations or values of some requirements may have to be rethought out. A satellite customer new to the business may be mystified by discussions with the payload manufacturer and need more knowledge to be able to get what he wants. Today it takes on the order of 10 years to "know" the payload. This book can accelerate learning.

The intended audience of this book is the following people who work with communications satellites:

- Payload systems engineers, at all stages in their careers
- Satellite-communications system designers and analysts
- Satellite proposal-writers
- Satellite customers

- Payload unit engineers
- Satellite control-center engineers, electrical and software
- Satellite bus engineers who need a payload reference now and then.

Prerequisite for full understanding of most of the chapters is knowledge of the Fourier transform and the duality of the time and frequency domains. However, without that almost of the book can still be understood.

My "love" for the satellite payload started from a rather early work experience. As a beginning engineer, armed with mathematical and some theoretical engineering knowledge, I attended a technical Space Shuttle meeting. One of the topics discussed was the screws on a unit. I felt disdain. Ten or 15 years later, for the first time being a systems engineer on a payload subsystem development, I learned that screws are fascinating—I saw the results of the unit's mechanical analysis. From then on I knew that things only seem boring when I do not know anything about them. Every bit of the satellite payload has to meet such stringent electrical and mechanical requirements that it has an intricate story behind it.

TERESA M. BRAUN

ACKNOWLEDGMENTS

I would like to express my deepest thanks to my husband, Walter Braun, who taught me communications theory on the job at LinCom Corp. in Los Angeles in the 1970s and 1980s and who has lovingly supported me and encouraged me in the writing of this book. Walter has a Ph.D. in Electrical Engineering with specialty in communications theory. I would also like to especially thank my Ph.D. advisor, Dr. Ezio Biglieri, for being so helpful and kind in the late 1980s when I was his graduate student and, a few years ago now, for his wonderful suggestion to write a book. I will always be grateful to all the wonderful engineers I have worked with over the years, especially Richard Hoffmeister and Dr. Charles Hendrix, who were instrumental in my career development. Almost all of the engineers I have worked with have been passionate about their work and willing to help others learn, and they have made mine such an interesting career. Of all the companies I have worked at, two stand out for having provided me limitless opportunities to do good work: Space Systems/Loral and William Lindsey's LinCom Corp. of 30 years ago. My career has spanned the time since the American equal-opportunity laws were being implemented at federal contractors, and I have gone from being an oddity in the engineering workplace to feeling at home among many women colleagues, in California, anyway. I wish to thank the colleagues who reviewed the book and provided corrections, suggestions, and explanations: my husband, who has read all of it and performed simulations; Richard Hoffmeister and Dr. Charles Hendrix, for reading most of it; Eddy Yee for Chapter 2; Gary Schennum for Chapter 3; Stephen Holme for the filter section of Chapter 4; James Sowers for Chapter 5; Dr. Messiah Khilla and Reinwald Gerhard for Chapter 6; Dr. Chak Chie for Chapters 8, 9, 11, and 13; and Dr. Ezio Biglieri for Chapter 10.

T. M. B.

ABOUT THE AUTHOR

Dr. Teresa M. Braun received the B.A. in Mathematics from the University of California, San Diego, in 1970; and from the University of California, Los Angeles, the M.A. in Mathematics in 1973, the M.S. in Systems Science in 1977, and the Ph.D. in Electrical Engineering in 1989 with dissertation on modulation and coding. She has also taken short courses on computer networks. She has been employed for 23 years in satellite communications and 3 years in satellite navigation. In California, from 1973 to 1976 she worked on GPS development at The Aerospace Corp.; from 1977 to 1986 on analysis and simulation of end-to-end satellite communications at LinCom Corp. (now LinQuest Corp.); from 1989 to 1997 in development of new payload technology in communications and navigation at Hughes Space & Communications (now Boeing Satellite Development Center); from 1997 to 1999 in development of new payload and ground-receiver technology and on a satellite constellation at Lockheed Martin's Western Development Laboratory (now part of LM's Management & Data Systems); and from 1999 to 2003 as a payload manager and department manager of payload systems analysis at Space Systems/Loral. After moving to Switzerland in 2003, she has worked in project management, supplier management, modem algorithm development, and payload analysis. She is still often in the US. She was *née* Thesken and also formerly named McKenzie. She has worked on NASA, defense, commercial, and ESA programs.

ABBREVIATIONS

$\circ$	Convolution
$*$	Complex conjugation
$\approx$	Is approximately equal to
$\triangleq$	Is defined as
$\ll$	Is much less than
$\gg$	Is much greater than
$\propto$	Is proportional to
$\varphi(f)$	Phase of transfer function in radians, a function of frequency f; also called "phase response"
ϕ	Azimuth angle in spherical coordinates
σ	Standard deviation
θ	Polar angle in spherical coordinates
$\theta(t)$	Signal phase in radians as a function of time t
8PSK	8-ary phase-shift-keying
16QAM	16-ary quadrature amplitude modulation
AC	Alternating current
ADC	Analog-to-digital converter
A/D	Analog-to-digital converter
$A(f)$	Signal-amplitude multiplication function of a filter as a function of frequency f; called for shorthand the "amplitude function" of the filter
AIAA	American Institute of Aeronautics and Astronautics, Inc.
ALC	Automatic level control
A-to-D	Analog-to-digital (converter)
AWGN	Additive white Gaussian noise
B	Bandwidth in Hz

BER	Bit error rate
BFN	Beam-forming network
B_L	Loop bandwidth (one-sided baseband) of phase-locked loop
BOL	Beginning of life
BPF	Bandpass filter
BPSK	Binary phase-shift-keying
C	Carrier or signal power in a given bandwidth
CAMP	Channel amplifier
CC	Conduction-cooled
coax	Coaxial cable
CONUS	Continental US
CP	Circularly polarized
CRO	Coaxial-resonator oscillator
CW	Continuous wave, that is, a sinewave
C/3IM	Ratio of carrier power to 3rd-order IMP power when nonlinearity's input is two equal-power carriers
D	Directivity; antenna aperture diameter
DAC	Digital-to-analog converter
D/A	Digital-to-analog converter
dB	Decibels
dBm	10 times log of milliwatts
DC	Direct current
D/C	Downconverter
DEMUX	Demultiplexer
DRC	Direct-radiation-cooled
DRO	Dielectric-resonator oscillator
D-to-A	Digital-to-analog (converter)
DUT	Device under test
DVB	Digital Video Broadcast
DVB-RCS	DVB–Return Channel Satellite
DVB-S	DVB–Satellite
$\boldsymbol{E}$	Electrical field vector
E_b	Energy per bit
EHF	Extremely high frequency (between 30 and 300 GHz)
E_s	Energy per modulation symbol
EIRP	Equivalent isotropically radiated power
EOC	Edge of coverage
EOL	End of life
EPC	Electronic power conditioner
E_s	Energy per modulation symbol
ESA	European Space Agency
ETSI	European Telecommunications Standards Institute
E–W	East-west; east and west
F	Flight
F	Noise figure

f	Frequency in Hz; focal length
f/D	Ratio of focal length to aperture diameter, for a paraboloidal reflector antenna
FDM	Frequency-dimension multiplexing
FGM	Fixed-gain mode
FIR	Finite-impulse response
FOV	Field of view
G	Gain
GaAs	Gallium arsenide
GEO	Geostationary orbit
$G(f)$	Filter gain, a function of frequency f; also called "gain response"
G/T_s	Antenna gain divided by system noise temperature
H	Horizontal linear polarization; hybrid coupler
$\boldsymbol{H}$	Magnetic field vector
HEO	Highly elliptical orbit
HF	Harmonic filter
$H(f)$	Filter transfer function, a function of f
HPA	High-power amplifier
HPF	High-pass filter
$h(t)$	Filter impulse response, a function of t
Hz	Hertz, unit of frequency
I	Interference power in a given bandwidth; in-phase component of baseband signal
IBO	Input backoff
IEEE	Institute of Electrical and Electronics Engineers, Inc.
IF	Intermediate frequency
Im	Function that takes the imaginary part of a complex number
IMP	Intermodulation product
IMUX	Input multiplexer
InP	Indium phosphide
ISI	Intersymbol interference
ITU	International Telecommunication Union
j	$\sqrt{-1}$
LCAMP	Linearizer-CAMP unit
LEO	Low earth orbit
LHCP	Left-hand circularly polarized
ln	Natural logarithm
LNA	Low-noise amplifier
LO	Local oscillator
log	Logarithm base 10
LP	Linearly polarized
LPF	Low-pass filter
LTWTA	Linearized TWTA
(L)CAMP	CAMP that may or may not contain the linearizer function
(L)TWTA	TWTA that may or may not be linearized

MAC	Medium-access control
MBA	Multibeam antenna
MEO	Medium earth orbit
MF	Matched filter
MLO	Master local oscillator
MPA	Multiport amplifier
MPEG	Moving Picture Experts Group
MPM	Microwave power module
MRO	Master reference oscillator
MUX	Multiplexer
N	Noise power in a given bandwidth
NASA	National Aeronautics and Space Administration
N_0	One-sided RF or IF power-spectral-density of noise
NF	Noise figure
NPR	Noise-[to-]power ratio. Ratio of IMP power in one slot to carrier power in a slot, when input to HPA is a large number of equal-power, equally spaced carriers with one missing in center slot
N–S	North-south; north and south
OBO	Output backoff
OBP	Onboard processor
OCXO	Oven-controlled crystal oscillator
OMUX	Output multiplexer
OQPSK	Offset QPSK
P	Long-term average of signal power. Average is over a time much, much longer than inverse of signal's noise bandwidth
P2dB	Amplifier's 2-dB compression point
$p(x)$	Probability density function of a random variable, a function of its values x
pdf	Probability density function
P_{in}	Power input to amplifier
$P_{\text{in sat}}$	Power input to amplifier that saturates amplifier
pk-pk	Peak-to-peak
PLL	Phase-locked loop
PM	Phase modulation
PN	Pseudo-noise
P_{op}	Amplifier's operating point
P_{out}	Power output by amplifier
$P_{\text{out sat}}$	Power output by amplifier when saturated
Pr	Probability
Preamp	Preamplifier
PS	Power supply
psd	Power spectral density
PSK	Phase-shift-keyed
$P(t)$	Instantaneous power of signal as a function of time t. Equal to signal's square magnitude averaged over carrier cycle

$p_X(x)$	Probability density function of the random variable X, a function of its values x
Q	Quadrature-phase component of baseband signal
QAM	Quadrature amplitude modulation
QPSK	Quaternary phase-shift-keying
R_b	Data bit rate of signal, in Hz
RC	Raised-cosine shape of signal spectrum
Re	Function that takes the real part of a complex number
RF	Radio frequency
RHCP	Right-hand circularly polarized
rms	Root mean square, namely the square root of the average squared value
RRC	Root-raised-cosine shape of pulse's Fourier transform
R_s	Modulation-symbol rate, in Hz
RSM	Regenerative Satellite Mesh
rss	Root-sum-square, namely the square root of a sum of squares
SEE	Single-event effects
SER	Modulation-symbol error rate
$s(t)$	Signal as a function of time t
$\mathcal{S}(f)$	Fourier transform of signal, especially when signal is a modulation pulse. A function of frequency f
$\mathcal{S}(f)$	Power spectral density of signal, a function of frequency f
SNR	Signal-to-noise ratio
SQPSK	Staggered QPSK
s/s	Samples per second
SSPA	Solid-state power amplifier
std dev	Standard deviation
t	Time in seconds
T	Repetition interval of modulation-symbol pulse
TCXO	Temperature-compensated crystal oscillator
TID	Total ionizing dose
TDM	Time-division multiplexing
TDRSS	Tracking and Data Relay Satellite System
T_s	System noise temperature
TWT	Traveling-wave tube
TWTA	Traveling-wave tube amplifier
U/C	Upconverter
V	Vertical linear polarization
VCO	Voltage-controlled oscillator
VNA	Vector network analyzer
VSWR	Voltage standing-wave ratio
WG	Waveguide
XPD	Cross-polarization discrimination

CHAPTER 1

INTRODUCTION

1.1 WHAT THIS BOOK IS ABOUT

This book is about the **payload** of communications satellites. The payload is what the satellite is *for*, namely the payload is what performs the functions desired of the satellite. In a nutshell, the payload is the communications antennas, receivers, and transmitters. The rest of the satellite, the bus, supports the payload by providing a structure, power, commanding, and telemetry, an appropriate thermal environment, radiation shielding, and attitude control.

The book has two parts, the first part having to do with the payload itself and as part of the satellite and the second part having to do with the payload as part of the end-to-end satellite communications system. Typically, the end-to-end system consists of a transmitter on the ground, the payload on orbit, and a receiver on the ground, as shown in Figure 1.1; this is what we will generally refer to. However, an end-to-end system could have another satellite as originating transmitter and/or as destination receiver, and virtually all of this book would still apply in this case; however, such a system has additional considerations which are not dealt with here.

The payload typically has on the order of a hundred pages of customer requirements. Analysis of these requirements is a large effort that starts even before the satellite-build program and goes on through to the end of the program. Every requirement must be met in a number of cases ranging from one to hundreds. The payload units, the payload layout on the spacecraft, and the payload's response to the on-orbit environment must be understood in detail. Initially the analysis is developed in enough detail that the customer sees he will get what he wants and the satellite manufacturer is confident of building it on schedule and making a profit.

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

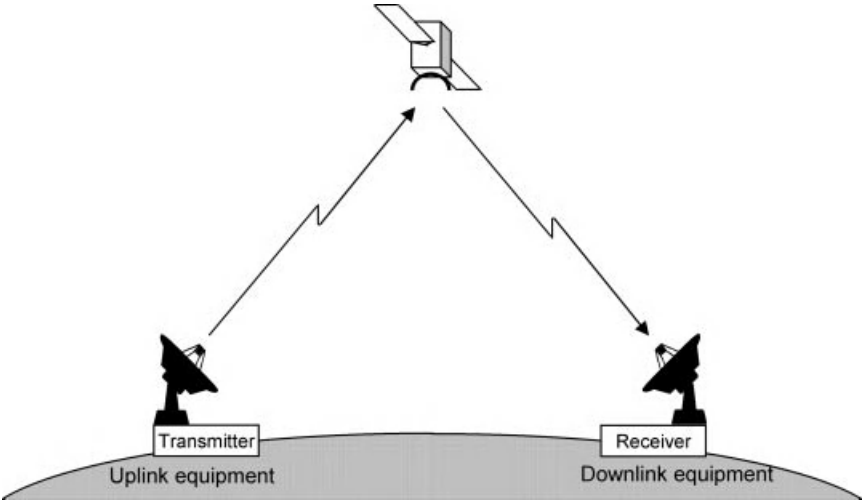


FIGURE 1.1 End-to-end satellite communications system, transmitter on ground through receiver on ground.

The first analysis phase during the program must show in detail that the customer's payload requirements are expected to be met, with margin, called the "risk margin." As units get built, better and better detailed numbers become available, and the analysis incorporates them and carries a decreasing risk margin. Finally, after testing, all the requirements are shown to hold and the customer accepts the satellite. This process is illustrated in Figure 1.2.

If a customer requirement cannot quite be met without a big negative impact on the satellite program, communications analysis of the end-to-end communications system can be decisive on how to proceed. It may persuade the payload engineer, the program manager, and the customer that the requirement can be loosened without degrading communications performance.

This book describes the payloads of commercial and governmental communications satellites from around the world that are on orbit today and providing a service. The governmental satellites included are space-administration (NASA and

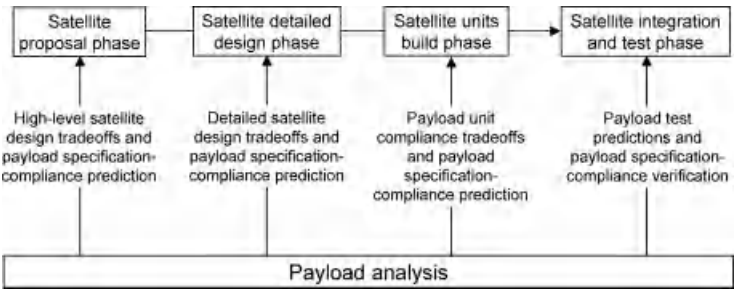


FIGURE 1.2 Contributions of payload analysis through all stages of satellite development.

ESA) and, where information is available, military. Experimental and proof-of-concept satellites are not included. Communications for command and telemetry of the spacecraft are not included. The carrier frequencies addressed are those in the GHz, chiefly from L-band (about 1.5 GHz) to Ka-band and to a lesser extent on to V-band (about 60 GHz).

1.2 PAYLOAD

This book is about the payload of communications satellites that are on orbit today. It talks about both kinds of communications payload, the **bent-pipe** and the **processing payload**. The bent-pipe is by far the most common payload type at present. A bent-pipe payload receives, filters, frequency-converts, switches, amplifies, and transmits signals via analog electronics. The processing payload has many elements in common with the bent-pipe payload but employs an onboard processor (OBP) to do more complex things than the bent-pipe can do. There are two classes of processing payload: (1) the **nonregenerative** class that has reconfigurable switching, digital filters at radio frequency (RF) (the uplink or downlink frequency) or at intermediate frequency (IF), and/or reconfigurable beams; and (2) the **regenerative** class that performs demodulation and remodulation at the least.

A payload must be described on at least two levels: (1) the basic structure of the **transponder**, which is the chain of **units** that supports a **channel's** path through the payload, where a unit is something that performs one major function of the transponder; and (2) the architecture of the overall payload. The properties of the elements of one level flow upward and/or downward to the elements of the other level. The whole thing must work together and do what it needs to do. And, oh, yes, "work" as a part of the overall spacecraft.

A payload processes signals in channels. (This definition of "channel" is not the same as in communications theory, where it means everything that happens to the signal between carrier modulation and data detection.) In a limited sense, a channel is a frequency band used to carry signal(s). The channel can be shared in one of the two ways, by time-division multiplexing (TDM) or frequency-division multiplexing (FDM), both illustrated in Figure 1.3. In TDM the various users of the channel share

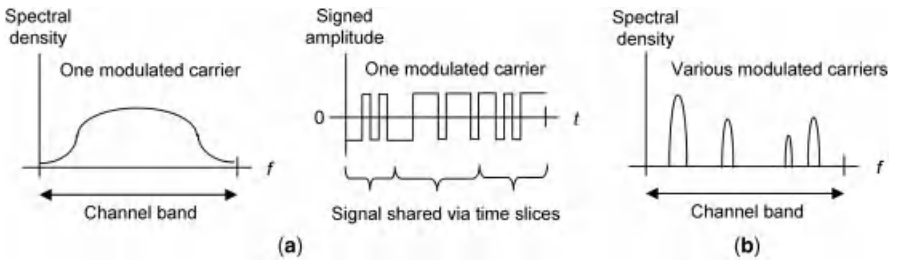


FIGURE 1.3 Two ways of channel-sharing, used on payload: (a) TDM, shown in both frequency domain and time domain and (b) FDM, shown in frequency domain.

in the time domain and one carrier is modulated, while in FDM the users share in the frequency domain and each modulates a carrier. In this book, usually we will say that a channel carries a signal, not signals, with the meaning that the signal is the totality of what the channel carries.

In the fullest sense, a channel is defined not just by a frequency band but also by the uplink antenna-**beam** and the downlink beam that carry it. A beam is a solid angle with origin at the payload (usually identified with its footprint or coverage on the earth) and a polarization. A transponder provides a processing path through the payload for such a channel. Most satellites have many transponders, and the transponders commonly share hardware. A particular transponder exists only as long as the payload switches are turned so as to allow the channel to be processed wholly; when the switches are turned other ways, the transponder as an entity is gone and some of its units may be incorporated into different transponders; the action may be later reversed or followed by a different such action. Typically most of the units in a transponder are shared by other transponders; other channels merge into this channel for a while, are separated out, and so on.

1.2.1 Bent-Pipe Payload Transponder

The first and lowest level of describing the payload is the transponder. The basic structure of a bent-pipe transponder is always pretty much the same; a simplified structure is shown in Figure 1.4. The transponder elements shown are fewer than in any actual transponder but they are enough to represent the major ones in any transponder. Note that the transponder is taken to include the receive and transmit antennas (which is not always the case in the literature). The signal is shown as travelling from left to right, and this convention is used throughout the book. The RF lines coming out of the receive antenna and going into the transmit antenna are shown as waveguide and the rest as coaxial cable, but other mixes are possible.

The bent-pipe transponder units are the following:

- Receive antenna, which receives the signal from the uplink coverage area to the exclusion of other areas.
- Preselect filter, which suppresses uplink interference.

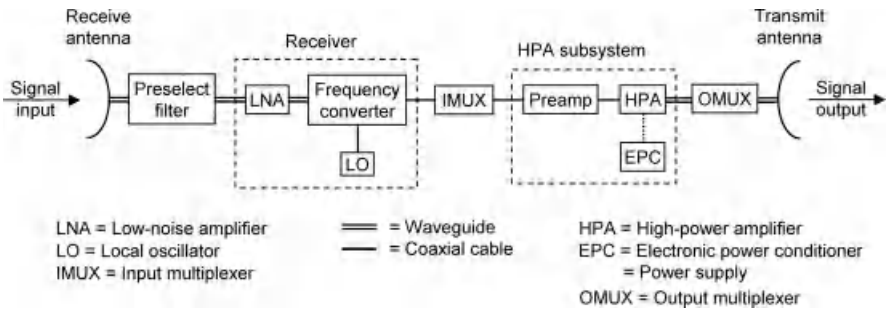


FIGURE 1.4 Simplified block diagram of one transponder of a bent-pipe payload.

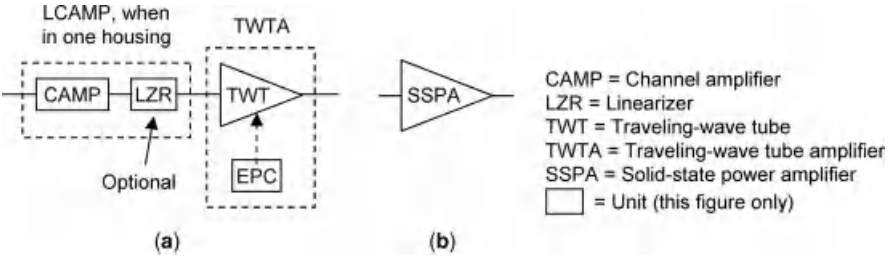


FIGURE 1.5 Two kinds of HPA subsystem: (a) TWTA subsystem and (b) SSPA unit.

- Low-noise amplifier (LNA), which boosts the received signal to a level where the noise added by the rest of the payload units will not cause serious degradation, and itself adds little noise.
- Frequency converter, usually a downconverter, to convert from the receive uplink frequency to the transmit downlink frequency. The LNA and the frequency converter are together sometimes called the "receiver."
- Input multiplexer (IMUX), a filter for separating out and/or combining channels.
- Preamplifier (preamp), which boosts the signal level to what the high-power amplifier needs to output the desired RF power.
- High-power amplifier (HPA), which boosts the signal level to what the transmit antenna needs to close the link.
- Electronic power conditioner (EPC) that provides power DC power to rest of the "HPA subsystem," namely the preamp and HPA.
- Output multiplexer (OMUX), a filter for combining and/or separating out channels.
- Transmit antenna, which transmits the signal to the downlink coverage area to the exclusion of other areas.

Each unit has a large number of variations. The one we mention now is that the HPA subsystem comes in two main varieties, one with a travelling-wave tube amplifier (TWTA) and the other with a solid-state amplifier (SSPA), as shown in Figure 1.5. The TWTA-type subsystem has three or four units as shown, including the channel amplifier (CAMP), which is the preamp. The SSPA usually incorporates all functions into just one unit.

1.2.2 Processing Payload

The nonregenerative processing payload has transponders such as those of the bent-pipe payload, aside from the OBP. However, the regenerative payload is very different. At its core is a signal processor. A simplified diagram of an overall regenerative payload is given in Figure 1.6. There may also be an upconverter bank between the modulator bank and the CAMP and HPA bank. Many of the units are unique to this kind of payload.

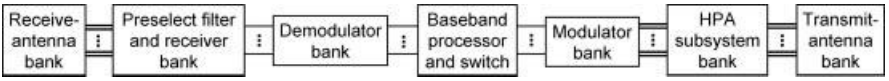


FIGURE 1.6 Simplified block diagram of a regenerative payload.

1.2.3 Overall Payload Architecture

The top level of payload structure is the overall payload architecture. The block drawing of this is the payload block diagram, which shows all the payload elements, namely units and payload-integration elements such as waveguide and coax pieces and switches. It also shows all the connections between elements. Drawing this is the first thing that the payload systems engineer does. An excellent reference for the variety in payload architectures is (Martin et al., 2007), which also provides high-level summaries of all current communications satellites and their businesses.

1.3 CONVENTIONS

The term **payload engineer** is sometimes used in the book, but the meaning is always the payload systems engineer, that is, the engineer responsible for making sure that the customer's payload requirements are met. This complex and rewarding job requires understanding and always keeping in mind the customer's written and unwritten requirements, negotiating with the payload-unit engineers on the allocation or flowdown of the customer requirements to unit level, working cooperatively with the unit engineers and bus engineers, being a main interface between the customer and the satellite program under the guidance of the program manager, keeping analysis at all times that shows all customer requirements will be met, formally accepting the units, and in some companies planning and supervising the payload test.

The frequency bands have various naming schemes worldwide, but in this book we will use the definitions adopted as a standard by the IEEE (2003) and shown in Table 1.1. Be aware that **K-band** is more and more often called "Ka-band" but not in this book. The reason for the increasing use of the misnomer is probably that

TABLE 1.1 IEEE Frequency Band Designations, L-Band to V-Band

Frequency (GHz)	Band
1–2	L
2–4	S
4–8	C
8–12	X
12–18	Ku
18–27	K
27–40	Ka
40–75	V

30 and 20 GHz are paired as up- and downlink spacecraft frequencies, respectively, and 30 GHz really is Ka-band, so calling the pair "Ka-band" is a shortcut.

A term is written in bold font where it is defined or at least described for the first time, if the term is used elsewhere in the book or is just important to know. A term is instead surrounded by parentheses where it is defined or described for the first time, if the term is not used elsewhere and is of lesser importance.

In this book, "t" means time and "f" means frequency. Other lowercase letters with "(t)" after them are time-domain functions, uppercase letters with "(f)" after them are frequency-domain functions, uppercase italic are Fourier transforms, and uppercase italic script are power spectral densities. The letters i, k, l, m, and n are integers. Temperature degrees are in Centigrade unless specified as being in Kelvin.

Theoretical definitions and properties are given with an emphasis on practicality and comprehendability. Rigor in stating all conditions of applicability is de-emphasized, because in real payload life the conditions are virtually always filled. References are given that contain the rigor, though. The reader who suspects he has a truly pathological case on his hands must look to those references.

1.4 BOOK SOURCES

Almost the entire book has come from publicly available sources, all in the English language. The sources in order of most-to-less frequently used are the IEEE online technical library Xplore, Google Scholar, and general Google, books, and the AIAA online technical library. Through Google, company catalogs, microwave encyclopedias, and patents were to be found. Wikipedia was a good source for background information whose reliability is not critical to a payload engineer. Some technology companies publish regularly and some scarcely publish. There must be equally interesting, innovative, and well performing technology that is not published in English that could not find its way into this book.

1.5 SUMMARY OF REST OF THE BOOK

The rest of the book is divided into two parts, Part I about the payload and Part II about the payload as part of the end-to-end communications system. Part I contains things that every payload engineer needs to know. Satellite customers, payload unit engineers curious about other parts of the payload, satellite bus engineers curious about the payload, and ground-station engineers would also be mainly interested in this material. Part II is additional material that people need, those who analyze or model the end-to-end communications system.

Chapter 2 is about the on-orbit environment that the payload sees and what happens to the payload because of it. The spacecraft orbit is a big driver of the spacecraft and payload layout. On-orbit effects are divided into thermal, aging, radiation, and attitude disturbances. What these do to the payload, after the spacecraft bus ameliorates them, is briefly summarized (detail is provided in the payload-unit chapters).

Chapters 3 through 6 describe payload units and payload-integration elements. Except for sometimes the IMUXes and OMUXes, these things are parts of both nonprocessing and processing payloads. Chapter 3 is on the antenna, Chapter 4 on the passive analog filter and the payload-integration elements, Chapter 5 on the LNA and the frequency converter, and Chapter 6 on the HPA subsystem. These chapters all present nearly the same categories of information, in some, the subset of payload architecture that has to do with the unit and, in all, the unit architecture, technology, performance, new flexible architectures, environmental considerations, and specification. Needed concepts are explained as they arise. Important terminology is defined. Some of the purposes of these chapters are to enable the payload engineer to understand product information, to know what to look for in a specification, to know what is a "regular" unit architecture and what is clever or risky, to satisfy his curiosity, and to partially appreciate and understand what a unit engineer is dealing with. In more detail, Chapter 3 presents concepts that apply to any antenna, then the various kinds of payload antenna, then autotrack. Chapter 4 starts by explaining impedance mismatch; then presents coax and waveguide; isolators, hybrids, and other integration elements; various kinds of passive analog filter; and last switches and the topic of redundancy. Chapter 5 explains the nonlinearity of the LNA and frequency converter; then presents the LNA; and finally the frequency converter including phase noise. Chapter 6 is on the HPA subsystem, first explaining HPA concepts; comparing the TWTA-based subsystem and the SSPA; then exploring both subsystem forms in details, including linearization.

Chapter 7 is different from the preceding chapters in that it explains the key communications parameters that are specified in the highest-level payload requirements document. It explains what these parameters are, what units and payload-integration elements contribute to them, how to approximate overall payload performance from units' performance (for predicting payload performance), and briefly how to test these parameters. It does not explain what the parameters mean to the communications signal, since this requires some knowledge of communications theory, which is provided in a later chapter.

Chapter 8 presents additional payload-level analyses, some of which are crucial to know and some of which may come in handy some time. The crucial ones include dealing with noise figure and keeping and maintaining performance-prediction budgets. The extra analyses include different ways to look at what the HPA does to the signal and an introduction to practical probability theory.

Chapter 9 is on the processing payload. The capabilities of current processing payloads are summarized. The technology specific to these payloads is described.

Part II starts with Chapter 10, which presents the principles of digital communications theory as applied to the satellite-based communications system. Because the theory topics are not separable, the presentation is in two rounds, first at a fundamental level and then at an intermediate level. No theorems are proved, and many drawings are provided.

Chapter 11 is on the communications link, that is, what comprises the signal paths besides the payload transponder and the ground station equipment chains. The first topic is end-to-end carrier-to-noise ratio. Next is what causes the signal

power on the link to vary, including atmospherics and then what establishes the noise power on the link and what makes it vary. An important topic is the three domains of interference. Last is an example of a link budget.

Chapter 12 is on how to model and optimize the performance of a multibeam downlink with atmospherics, using probability theory. The variations and uncertainties over the payload life of signal, interference, and noise are included. An example is worked through.

Chapter 13 is the last chapter and is about modeling the end-to-end communications system with focus on payload performance. The two types of modeling are software simulation (e.g., with Matlab) and hardware emulation (with a testbed). The essential basics of simulation, including how to model the HPA, are presented. Some easy-to-overlook issues in emulating the links are addressed.

REFERENCES

Martin DH, Anderson PR, and Bartamian L (2007). *Communications Satellites*, 5th ed., El Segundo, CA; Reston, VA: The Aerospace Press; American Institute of Aeronautics and Astronautics, Inc.

IEEE Std 521-2002 (2003). *Standard Letter Designations for Radar-Frequency Bands*, Piscataway, NJ: IEEE.

PART I

PAYLOAD

CHAPTER 2

PAYLOAD'S ON-ORBIT ENVIRONMENT

This chapter describes the payload's on-orbit environment and the general effects of the environment on the payload. How the payload units are affected in detail is addressed in Chapters 3–6. The sections in this chapter are as follows:

- *Section 2.1*: What determines the on-orbit environment, namely the orbit, the spacecraft's general layout and orientation, and the payload's general layout on the spacecraft bus
- *Section 2.2*: On-orbit environment itself and how the spacecraft bus mitigates it for the payload. General aspects of the environment are thermal, aging, radiation, and attitude disturbances
- *Section 2.3*: General effects of the mitigated environment on the payload.

2.1 WHAT DETERMINES ENVIRONMENT

The payload's on-orbit environment is determined, at the highest level, by the spacecraft orbit, the spacecraft's general shape and configuration, the spacecraft's orientation on orbit, and how the payload is laid out in the spacecraft.

2.1.1 Orbit

Communications satellites can be found in various **orbits**. The satellite revolves about the center of the earth in either a nominal circle or a nominal ellipse. The

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

nominal shape is characterized by the **eccentricity**, which is zero for a circle and positive for an ellipse. The circle or ellipse is only nominally the motion because it is perturbed by the nonspherical shape of the earth, local variations in the earth's density, the gravitational pull of other heavenly bodies, and other effects. The satellite's path lies in a plane that contains the center of the earth. This plane makes an angle with the earth's equatorial plane, called the **inclination angle**, which stays nominally constant. The highest latitude that the satellite reaches in its orbit is equal to the inclination angle. A communications satellite moves generally eastward in its orbit plane, similarly to the earth, which moves exactly eastward. The nominal motion of the satellite is independent of the movement, that is, rotation, of the earth below, but the partial synchrony of the two means longer periods of time where communications is possible. (Other kinds of satellites may move generally west to east when their mission does not require good contact with the earth below.) Another aspect of the orbit is its period, which along with eccentricity determines its altitude profile. Yet another aspect, although dependent on the others, is the speed variation over the orbit, and a satellite in elliptical orbit will move faster while it is near the earth and more slowly while it is far away. The various aspects of the satellite's orbit and, if there is more than one satellite in the constellation, of the other satellites' orbits are chosen by the system designer to serve a particular communications purpose well. If an equatorial orbit with period equal to the earth's will not serve the purpose well then there is probably more than one good solution and a complex tradeoff must be made. Further description of orbits, orbit trade-offs, and orbit perturbations is given in Evans (1999).

The three types of orbits in use today for communications satellites are the following:

- *Geostationary Orbit (GEO)*: The altitude above the earth is about 36,000 km. The satellite revolves in a circle in the equatorial plane at the same speed at which the earth rotates, so the satellite is very nearly stationary with respect to the earth underneath. Over the day the subsatellite point on the earth traces a small figure-eight, where the crossing-point of the figure-eight is on the equator. The largest deviation north or south of this figure-eight is less than 1° for most communications satellites (Morgan and Gordon, 1989) and equal to the orbit plane's inclination. An exception is the Inmarsat-4 series of satellites, whose inclination starts at 3°, goes to 0°, then back to 3° over the design life (Martin et al., 2007). The communications' coverage area of a GEO satellite is in the mid-latitudes. An uncommon variation on this orbit is the **geosynchronous orbit**, which is a geostationary orbit gone bad. It is not an orbit of choice but happens when the satellite has run out of station-keeping fuel to keep it geostationary. The satellite may still be partially used this late in its life but not as a primary satellite for any service.
- *Low-Earth Orbit (LEO)*: The altitude of the orbit is about 1000 km, and its orbital period is a little over an hour and a half. The orbit is circular.

- *Highly Elliptical Orbit (HEO)*: HEO in which the apogee's altitude is much higher than the perigee's. The satellite moves very fast near perigee and relatively slowly near apogee. HEOs generally have an inclination of 63.4° because there the orbit is the most stable for the least station-keeping fuel. One type of HEO is the **Molniya orbit**, which has a period of about 12 h. The exact altitudes of apogee and perigee are not defined but for the original Molniya orbit of the former USSR the altitudes were about 39,000 and 1,000 km, respectively. Another type of HEO is the **tundra orbit**, which has a period of about 24 h. For both these orbits, when the satellite is near apogee it seems to hang high in the sky in about the same place for 2/3 of the orbital period (Maral and Bousquet, 2002).

A further orbit, not currently in use but planned to be in 2013 by the O3b constellation (O3b, 2010), is the following:

- *Medium-Earth Orbit (MEO)*: The altitude of the orbit is between 5,000 and 12,000 km (Elbert, 2004).

The orbits for some current communications satellite systems that serve different purposes are the following:

- *Most Communications Satellites*: GEO.
- *Globalstar Satellite System*: LEOs at altitude of 1400 km, 6 satellites in each of 8 orbit planes, 52° inclination. System application is global mobile telephone and low-rate data relay (Martin et al., 2007).
- *Iridium Satellite System*: LEOs at mean altitude of 780 km, 11 satellites in each of 6 orbit planes, 86° inclination. System application is global mobile telephone and low-rate data relay (Martin et al., 2007).
- *Sirius Radio Satellite System*: HEOs in tundra orbit, perigee at 24,500 km and apogee at 47,100 km, 3 satellites each in its own orbit plane, all satellites having the same ground trace, 63.4° inclination. System application is mobile radio broadcast to CONUS (Briskman and Prevaux, 2001).
- *National Reconnaissance Office's Satellite Data System, 3rd Generation*: HEOs in Molniya orbit, USA 198 launched in 2007 having perigee at 515 km, apogee at 47,100 km, and 62.6° inclination. System application is suspected to be relay of reconnaissance data (Wade, 2010).

2.1.2 GEO Spacecraft's General Layout and Orientation

By far the most communications satellites today are three-axis stabilized spacecraft. A GEO satellite is usually a rectangular box with external attachments. The earth-facing side, the **earth deck**, is often square, otherwise rectangular. It has antennas mounted on it, usually more than one. There are four rectangular sides attached to the earth deck, namely the **north, south, east,** and **west panels**. Many spacecraft

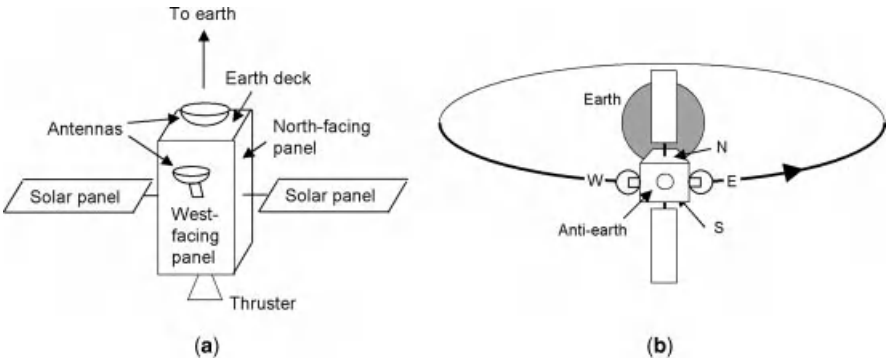


FIGURE 2.1 GEO spacecraft: (a) typical overall layout and (b) orientation on orbit.

have antennas attached to the east and west panels. Most are attached by short booms but some are attached by very long booms, for example, on the TerreStar spacecraft, because the antenna is so large, 18 m across. The solar panels are attached to the north and south panels. The face opposite the earth deck, the **anti-earth deck**, has thrusters on it. This typical overall layout of a GEO spacecraft is shown in Figure 2.1.

Figure 2.2a shows the DirecTV-7S spacecraft, launched in 2004, which is a rather typical GEO spacecraft. It has a few reflectors on the earth deck. A very unusual GEO is the Milstar spacecraft, shown in Figure 2.2b. The bus for this one is the rectangular section in the middle that is not in a direct line with the rest of the long-axis structure. Presumably the structure is so long because of the very large number of antennas that it has to accommodate.

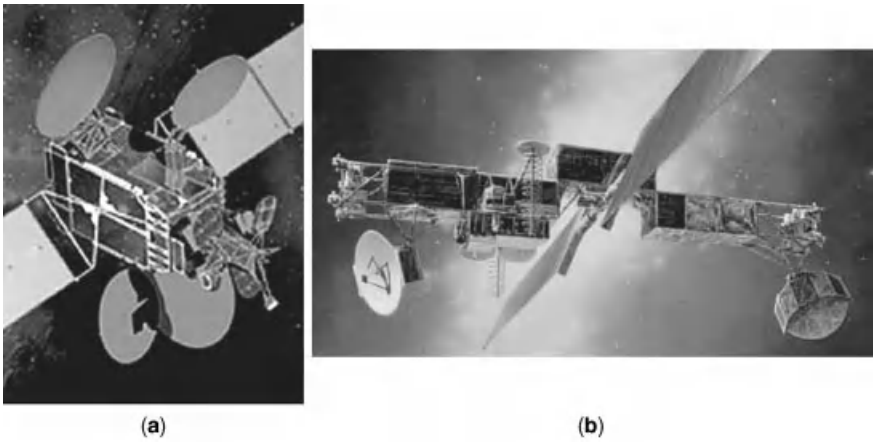


FIGURE 2.2 GEOs: (a) DirecTV-7S (image courtesy of Space Systems/Loral) and (b) Milstar (image courtesy of US Air Force).

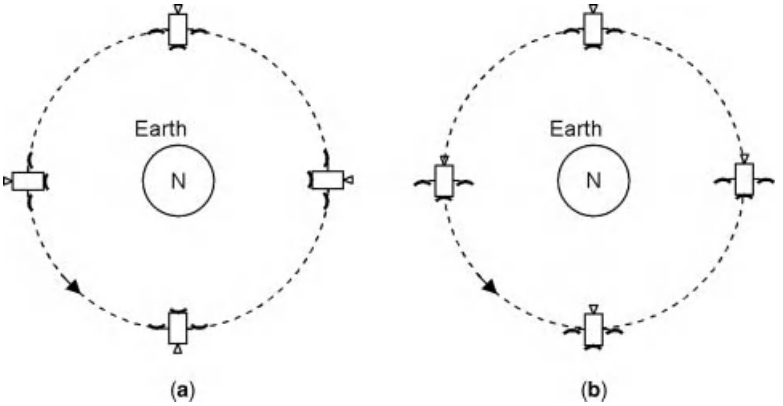


FIGURE 2.3 Orientation of equatorially orbiting spacecraft, over one orbit: (a) with attitude control and (b) without attitude control.

The GEO spacecraft's orientation on orbit is determined by two constraints. One is that the earth deck must point down to the earth. The other is that the solar arrays need to have sun all the time except when there is no sun, during eclipse. All GEO communications satellites have their solar arrays attached to the north and south panels, and the solar arrays rotate about a north–south axis. In this way, the arrays are always nearly perpendicular to the sun.

The satellite must have its attitude (i.e., the direction in which it faces) controlled, to keep the earth deck facing the earth as the satellite moves through its orbit. If the satellite had no attitude control, it would rotate about the earth always facing the same direction with respect to the stars, as shown in Figure 2.3. The satellite goes around the earth once per orbit, that is, once per day. So the attitude of the spacecraft is controlled to make it rotate about its north–south axis once per orbit.

2.1.3 GEO Spacecraft's Payload Configuration

Now we look inside the spacecraft. Let us imagine it sliced and opened out as you might do with a cardboard box, so we can see the inner surfaces. The anti-earth deck has been thrown away because typically it contains no payload equipment. Two examples of the payload hardware layout on the spacecraft are shown in Figure 2.4. Ideally, all the payload equipment would be right at the top (in this orientation) of the panels to minimize losses from waveguide and coaxial cable runs, but there is too much payload hardware for this, so a compromise must be made. One driver of the payload configuration in Figure 2.4a is that the receive antenna is mounted on the earth deck (not shown). The LNAs must be close to the receive antenna, because long transmission-line runs to the LNAs would degrade the SNR unacceptably. The frequency downconverters should be close to the LNAs so that the long cable run to the IMUXes is at the lower frequency, where the cable loss is less. Now, TWTAs

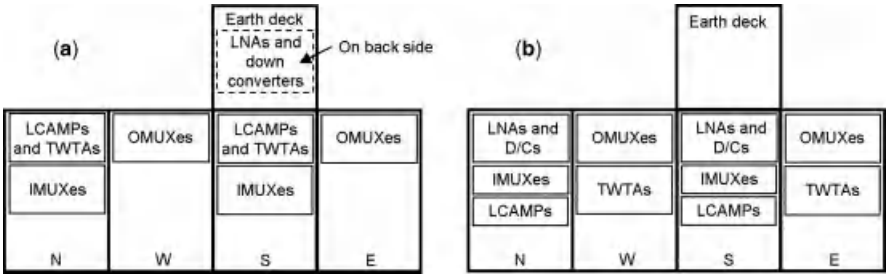


FIGURE 2.4 GEO spacecraft opened out to reveal inside two example configurations of bent-pipe payloads with TWTAs.

consume the most DC power of any payload equipment, and the RF power out of the TWTAs cannot be wasted, so the TWTAs must be placed close to the OMUXes and the OMUXes must be placed close to the transmit antennas. In this example, the transmit antennas (not shown) could be on the earth deck or the east and west panels. Finally, the IMUXes have to go on the north and south panels, where they will be cooler than on the east and west, so they go below the TWTA subsystems. The second example in Figure 2.4b has the LNAs and downconverters moved from the earth deck to the north and south panels because the receive antennas are on those panels (not shown). The moving of the TWTAs from the north and south panels to the east and west is the product of recent improved flexibility in the spacecraft thermal subsystem (Section 2.2.1.6). There are many transmission lines and minor passive elements, which are not shown here. Another driver of the configuration is the need to minimize the total amount of coaxial cable and waveguide, for weight and space. Units or banks of units that have many RF lines in between would be preferably placed close together. The payload equipment cannot take the entire panels because bus equipment has to go on them, too.

2.1.4 Non-GEO Spacecraft Considerations

Non-GEO spacecraft are freed from the constraints imposed on a GEO but have different constraints.

Before we can describe how these spacecraft are oriented, we must define the **body axes** of a satellite. They are defined as if the spacecraft were an aircraft flying along (see Figure 2.5). The aircraft's pitch axis is the axis that goes through the wings. To make this apply to a satellite, we associate the pitch axis with the axis that goes through the solar panels, the yaw axis with the direction perpendicular to the earth deck, and the roll axis the one that completes the set. If the satellite were in GEO, the pitch axis would be north–south, and the solar panels would be aligned with and rotate about this axis.

A Globalstar-2 LEO spacecraft is pictured in Figure 2.6a. The first six were launched in October 2010 (the Globalstar-1 spacecraft are on orbit already and have a similar shape except for the antennas). The earth deck is the bottom of the

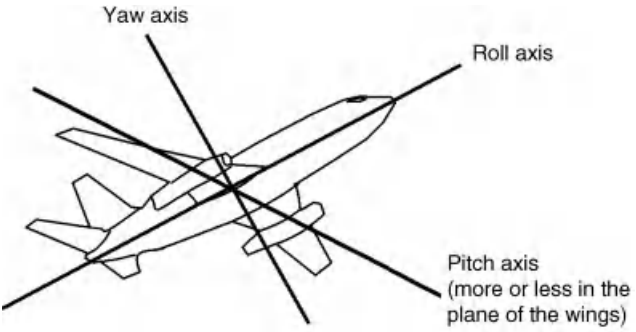


FIGURE 2.5 Body axes of aircraft.

spacecraft in the picture. It is rectangular instead of square like on most GEO satellites. The three large objects on it are a collection of phased array antennas. On the two of these objects that are conical, each panel is a phased array. All the panels do not point down as they would on a GEO spacecraft (see Section 3.5.3 for a discussion of the antennas). The rest of the spacecraft is shaped somewhat like a GEO satellite, but the side panels are small, so the ratio of earth-deck surface area (where as many antennas as possible are put) to side-panels surface area (where repeater electronics are placed, inside) is different.

The orbit's inclination is not near zero as for a GEO but is 52°, so if the solar panels could only be moved by the same means as on a GEO they would often be

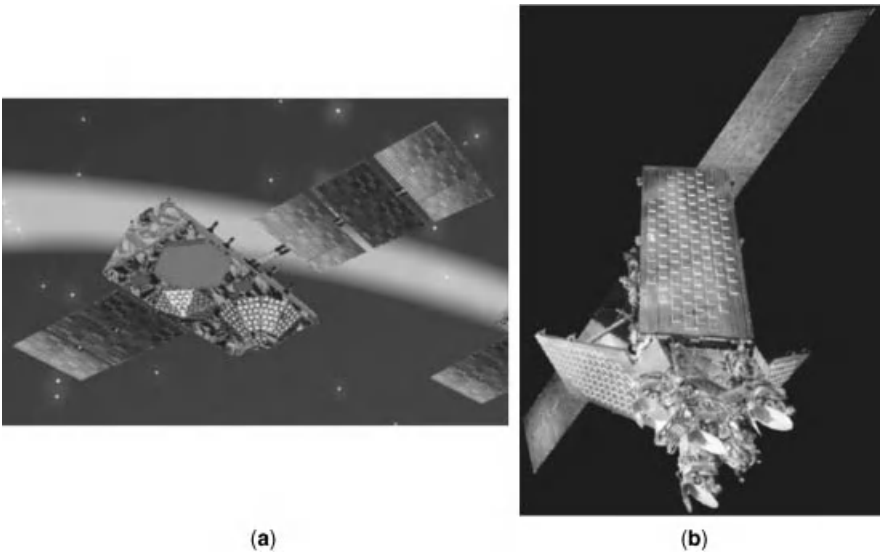


FIGURE 2.6 LEOs: (a) Globalstar-2 (E. Briot © Thales Alenia Space) and (b) Iridium (photo by Eric Long, National Air and Space Museum, Smithsonian Institution, SI 2006-935).

away from the sun. The solar panels do rotate about the pitch axis as on a GEO but additionally the satellite rotates about its yaw axis (Rodden et al., 1998). This is called "yaw steering."

An Iridium LEO spacecraft is pictured in Figure 2.6b. They were launched from 1991 to 2002. The earth deck, with the reflector antennas on it, is triangular. Attached to the earth deck and flaring away from it are three phased-array antennas (see Section 3.4.3 for a discussion on them). These antennas are not parallel to the earth as on a GEO satellite.

The Iridium satellites orbit with an inclination of 86°, almost a polar orbit. The solar panels have the usual one axis of rotation (JPL, 1999). The spacecraft does not yaw-steer but keeps a particular one of its three phased arrays always pointed forward (Mansfield, 2008). So it must be that the solar arrays are, as usual, along the satellite's pitch axis but in this case are oriented almost east–west.

The Sirius (US) HEO spacecraft were launched in 2000 (the "US" in parentheses is to distinguish it from the spacecraft of the same name owned by SES Astra of Luxembourg). They look like GEO spacecraft. This makes sense, since during most of the time the spacecraft is transmitting, it is far above the earth (higher than a GEO spacecraft) and its coverage area subtends a small angle below. So the antennas need to point basically down.

The satellites orbit with an inclination of 63°. A similar solution for pointing the solar arrays is used as for the Globalstar spacecraft, but the yaw steering is unused for two 1.5-month periods during the year (Briskman and Prevaux, 2001).

2.2 ON-ORBIT ENVIRONMENT

In this section and the ones that follow in this chapter, we deal only with GEO spacecraft.

2.2.1 Thermal

2.2.1.1 Cold of Space The temperature of space, without sunlight, is 3°K or −270°C. Payload electronics cannot operate at a temperature anywhere near this low, and even if they could it would be infeasible to perform satellite-level thermal testing down to this temperature.

2.2.1.2 Heat from Spacecraft Electronics The spacecraft electronics produce heat since no device is 100% power-efficient. This is the basic source of the heat that keeps the electronics warm. The active units have an operating temperature range of something like −20 to +65°C. (Note that this range is approximately centered around room temperature—this simplifies testing.)

2.2.1.3 Changing Direction of Sun from Satellite The changing direction of the sun from the satellite has two effects: one on the repeater and the other on reflector antennas.

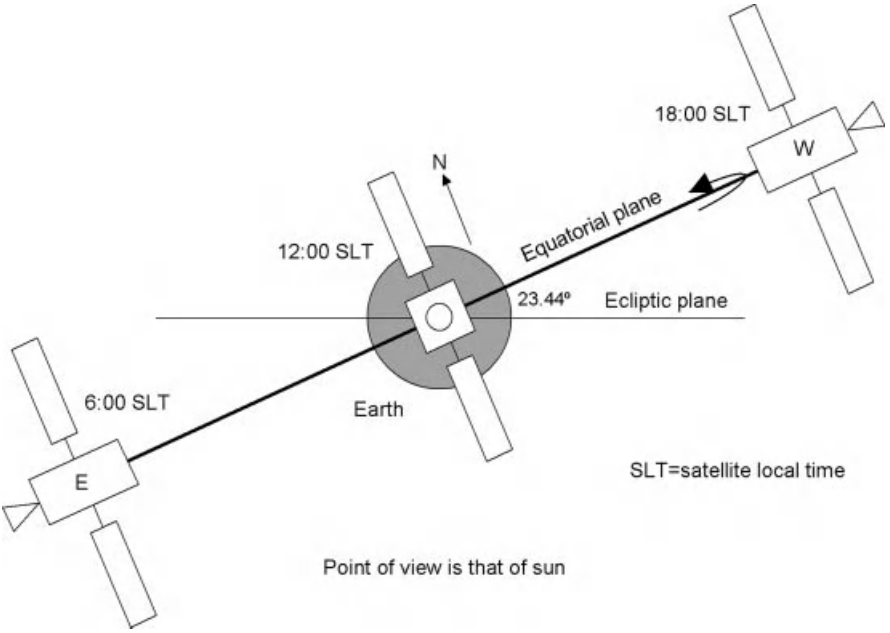


FIGURE 2.7 GEO satellite's diurnal motion at equinox.

The orientation of the satellite over the day and the course of the year determines how much and how long each of the satellite's faces will be in the sun. When a face is in the sun it is warmer than in the shade, and the more nearly perpendicular the face or panel is to the sun the warmer it is.

For a GEO satellite, the variation of the sun's direction over the day and year are illustrated in Figures 2.7, 2.8 and 2.9. What is illustrated is how the satellite looks

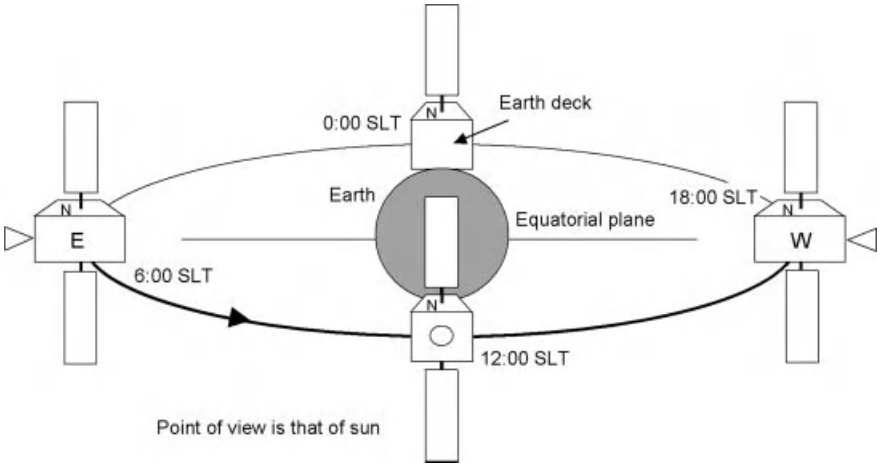


FIGURE 2.8 GEO's diurnal motion at northern hemisphere's summer solstice.

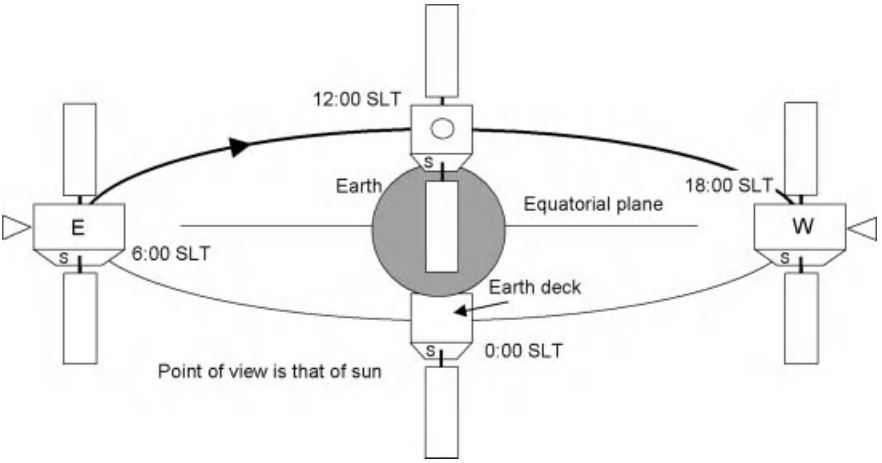


FIGURE 2.9 GEO's diurnal motion at northern hemisphere's winter solstice.

from the sun's viewpoint at equinox and at the northern hemisphere's summer and winter solstices, respectively, at four times of the day. At equinox, no sun is on either N–S panel, so there is minimal diurnal variation for these panels. For part of the day, sun is on the east panel and none is on the west, and for the other half the day sun is on the west panel and none is on the east, so these panels have a significant diurnal variation. At summer solstice, the north panel gets the same amount of sun all day long and the south panel gets none, so diurnal variation is minimal. The experience of the E–W panels at summer solstice is about the same as at equinox, although the amount of sunshine falling on them is a little less because of the 23.44° tilt of the orbit plane relative to the line between sun and earth. The same is true of the earth deck. Winter solstice is similar to summer solstice except that north and south have switched places. In summary, the N–S panels never receive much sun, while the E–W panels go from full sun to full shade and back over the course of a day. The N–S panels are on average cooler than the E–W panels and see a much smaller temperature swing.

The changing sun angle also affects the antennas. A reflector has a different amount of surface distortion depending on its temperature (how much sun is on it).

2.2.1.4 Changing Distance of Sun from Earth The earth's orbit about the sun is slightly elliptical, with an eccentricity of 0.0167. This causes a $\pm 3\%$ variation in sunlight strength reaching the earth-orbiting satellite (Morgan and Gordon, 1989). The difference this makes to the payload is only in the definition of testing temperatures: "cold" is defined for no sun and "hot" for full sun at minimum distance. This influence will not be discussed further.

2.2.1.5 Eclipse Eclipse is when the satellite is out of the sunlight, blocked by the earth. For a GEO satellite, the time in eclipse per day varies with the time of the

year. The eclipse season is 3 weeks before and 3 weeks after the equinoxes. There are no eclipses outside of these two periods. At an equinox, eclipse lasts about 68 min a day, and a week and a half before or after, it lasts about 60 min (Morgan and Gordon, 1989).

2.2.1.6 Mitigation by Bus Thermal-Control Subsystem The bus's thermal-control subsystem provides the right temperature ranges for the various elements of the payload. It carefully balances heat removal with heat retention, so the payload equipment becomes neither too hot nor too cold. The whole temperature control of the payload and the rest of the spacecraft is a complex design problem that the thermal engineer attacks iteratively, with the payload engineer in the loop if need be. In the rare cases when the subsystem cannot provide the desired temperature ranges for the payload elements, the payload has to be partially redesigned.

Each payload element that must dissipate heat has a **baseplate** as its bottom. The element is built so that the unit's heat flows to the baseplate, which gets attached to a spacecraft mounting plate. The spacecraft's thermal-control subsystem is designed to ensure that the mounting plate will be kept within a particular range of temperatures. The unit engineer designs the unit so that when its baseplate is in contact with a mounting plate at the maximum and minimum temperatures, every component in the unit will be within its required temperature range. If he cannot do that, the spacecraft thermal engineer has to adjust the thermal-control system to provide a better mounting-plate temperature range.

The mounting surfaces for the payload electronics, aside from the antennas and associated waveguide, are several, shown with the very thick dashed lines in Figure 2.10, the four side panels' inside surfaces, internal equipment panels, and the outer surface of the earth deck. Figure 2.10a is what the spacecraft looks like

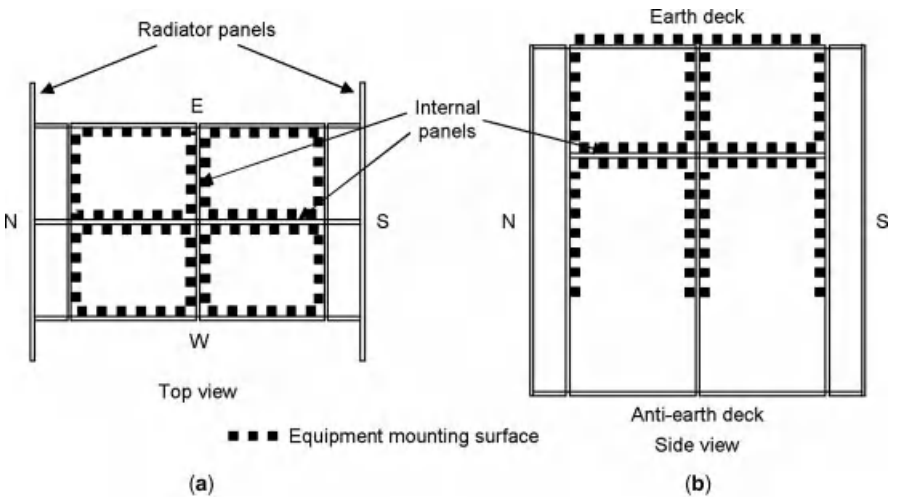


FIGURE 2.10 Typical spacecraft configuration of payload-equipment surfaces (after Yee 2002).

if the earth and anti-earth decks are removed so one can look into the spacecraft. Figure 2.10b shows the spacecraft from the side without the east and west panels. The interior panels are optional, as are their particular width and length. (Spacecraft bus equipment is also mounted to these surfaces and to the lower part of the panels shown without thick dashed lines.)

Embedded in some of the equipment mounting panels is a matrix of **heat pipes**, which carry waste heat away from the spacecraft mounting plates. A heat pipe is a passive, self-contained device that evaporates a working liquid at its hot end, carries the gas to the cooler end where it condenses, and by capillary or wicking action carries the liquid back to the hot end (Cullimore, 1992). Other heat pipes carry the heat from the equipment panels to the **radiator panels**. In the radiator panels there are further heat pipes. The outer radiator surface is of "optical solar reflector," which reflects optical sunlight while behaving like a black-body radiator for thermal radiation (Hosick, 2000). As shown in Figure 2.10, radiator panels typically cover the outside of the entire north and south equipment panels (Watts, 1989). In the last decade, there have been breakthroughs in the use of heat pipes, expanding the range of possibilities of the payload configured on the spacecraft. The capacity of the thermal subsystem has been increased, so that the payload can use and transmit more power. Unit temperature swings are reduced, decreasing the payload-performance variation. Some important breakthroughs are as follows:

- Potential for thermal coupling of north and south radiator panels by heat pipes, enabling the equipment on north and south to have more constant temperatures (Watts, 1989).
- Potential for thermal coupling of east and west panels, reducing the daily temperature swing of those panels and allowing a wider range of equipment to be mounted there, such as the TWTAs, switches, waveguides, OMUXes, power combiners, and loads (Hosick, 2000).
- Potential for thermal coupling of earth deck to north and south panels, reported in Jansson (2011).
- Potential for thermal coupling of north and south panels to east and west, to dump heat from north and south to east or west (Jondeau et al., 2010).
- Introduction of loop heat pipes, which have much higher capacity and can transport heat a longer distance than regular heat pipes. Loop heat pipes are used to connect the equipment panels to the radiator panels and even to multiple radiator panels (Yee, 2002). Introduction of bidirectional loop heat pipes, which can transfer heat between equipment panels and also between radiators (Yee, 2003).

Another part of the thermal subsystem is spacecraft surface finishes (e.g., paints and tapes) (Yee, 2011).

The last part of the subsystem is the **thermal blanket**, which covers the outside of the spacecraft except for antennas, thrusters, and radiators. It is multilayer to ensure that virtually no heat generated by the spacecraft electronics escapes (Wertz and Larson, 1999). It is shiny on the outside to reflect sunlight.

2.2.2 Bus and Payload Aging

An active payload unit ages as it is operated (and does not age when it is not operated).

The spacecraft's thermal subsystem ages, in which the radiators' outer surfaces and other spacecraft surface finish age, reducing their capacity to reflect sunlight or radiate heat (Yee, 2011). The payload equipment on heat pipes becomes warmer as the years go by unless heaters are employed in the early years for compensation.

2.2.3 Radiation

Radiation is transfer of energy by means of a particle (including photons) at X-ray frequencies or higher (ECSS, 2008b). Primary radiation comes from outside the spacecraft, and secondary radiation comes from the interaction of primary radiation with the spacecraft (ECSS, 2008a). Primary radiation originates from the galaxy all around and specifically from the sun. Cosmic rays, which are actually particles, come from the galaxy and are 89% protons, 10% helium nuclei, and 1% electrons (Wikipedia, 2010a). The solar wind ultimately provides the rest of the primary radiation. It is a continuous stream of charged particles, mostly electrons and protons, ejected from the sun. It is stronger during the more active part of the approximately 11 year solar cycle. Sometimes during especially high solar activity, bursts of solar energetic particles last from a few hours to several days. These particles are protons, electrons, and heavy ions of much higher than normal energy. The stronger the solar wind is, the more galactic cosmic-ray particles it sweeps away from the earth (ECSS, 2008a).

High-energy electrons and protons can harm some of the payload electronics. Spacecraft structural enclosures are typically made of aluminum, a few millimeters thick, as shielding. In order to penetrate 2 mm, an electron must have energy greater than about 1 MeV and a proton greater than about 20 MeV (Daly et al., 1996).

The earth is protected from the particles by its magnetic field, which deflects some galactic cosmic-ray particles (Wikipedia, 2010a) and most solar particles (Wikipedia, 2010b) and traps other particles. To first order, the earth's magnetic field is a magnetic dipole tilted about 11° off the earth's rotation axis with its center about 500 km to the north of the equator (Wikipedia, 2011). The places where the imaginary dipole goes through the earth's surface are the geomagnetic poles. The trapped particles form the Van Allen radiation belts. The belts for electrons of greater than 1 MeV are shown in Figure 2.11a. A Molniya orbit would go through the edge of the belts. On the geomagnetic equatorial plane, the strongest part of the electron belts is at about 4.3 earth radii from the center of the earth. Figure 2.11b shows the belts for protons of energy greater than 10 MeV. These belts are smaller than the electron belts. The strongest part of the belt in the geomagnetic equatorial plane is at about 1.75 earth radii from the center of the earth. As can be seen in Figure 2.11b but better in Figure 2.11c, the proton belts come especially close to the earth's surface in the southern hemisphere, making the South Atlantic anomaly. The center of the anomaly is at about 35° W longitude (Daly et al., 1996), which is

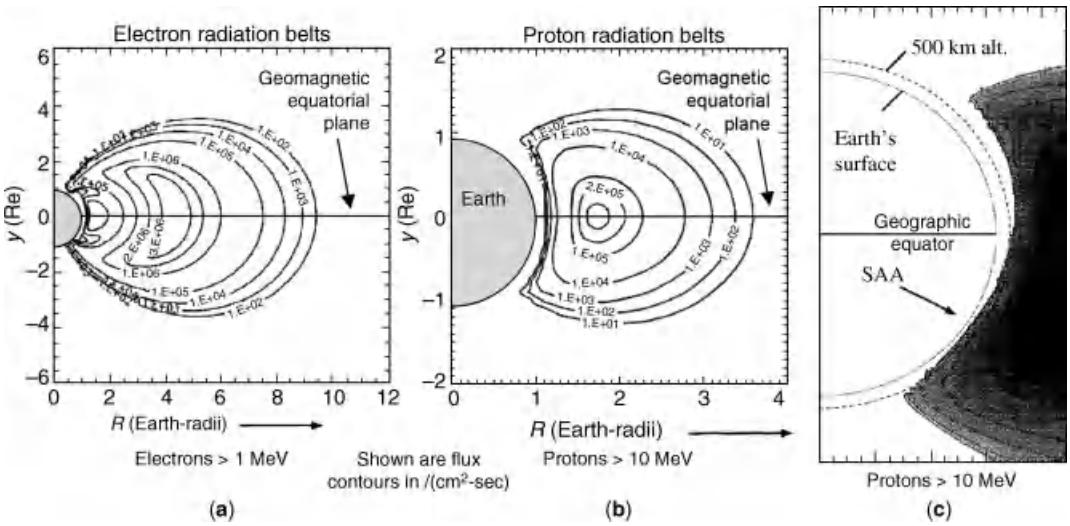


FIGURE 2.11 Van Allen radiation belts: (a) electron belts; (b) proton belts; (c) South Atlantic anomaly is shown in earth cut longitudinally at 35° W. (©1996 IEEE. Reprinted, with permission, from Daly et al. (1996)).

about the longitude of the eastern edge of South America. The anomaly can impinge on a LEO orbit. This is believed to be the cause of the degradation of the S-band power amplifiers on the first-generation Globalstar satellites (Selding, 2007).

2.2.4 Spacecraft Attitude Disturbances

The most important attitude disturbances external to the spacecraft are gravity gradient, solar radiation pressure, magnetic-field torques, and, for LEOs, aerodynamics (Wertz and Larson, 1999). The principal disturbances internal to the spacecraft are uncertainty in the center of gravity, thruster misalignment, mismatch of thruster outputs, rotating machinery, sloshing of liquids such as fuel, and dynamics of flexible bodies such as antennas, booms, and solar panels (Wertz and Larson, 1999). The worst disturbance in attitude is the brief burning of the spacecraft thrusters for an orbit correction or momentum dump.

The spacecraft is designed to keep its attitude under control well enough so that the antenna pointing is good enough to support the payload's mission.

2.3 GENERAL EFFECTS OF ENVIRONMENT ON PAYLOAD

The on-orbit environment affects the payload in the ways summarized in Table 2.1. In the subsections below we discuss these and what some payload units do to further mitigate the environment. How the general effects translate into performance effects for the various payload units will be discussed in Chapters 3–6.

2.3.1 Temperature-Variation Effects

Temperature variations of the payload units come on three time scales for any orbit: once per orbit, yearly, and over life. For a GEO satellite the variation once per orbit means once per day. For payload electronics on heat pipes, the once-per-orbit and yearly variations are not simply periodic. As the thermal subsystem ages, the once-per-orbit and the yearly variations may grow in size. The growth is between linear and logarithmic (Yee, 2011). In addition, the unit's average temperature drifts

TABLE 2.1 General Effects of On-Orbit Influences

Environment Aspect	General Payload Effects
Thermal with the exception of reflector front going in and out of sun	Temperature range of thermal interface (mounting plate) that spacecraft presents to unit
Radiation	Irreversible changes in RF elements; reversible changes in digital
Aging	Increased temperature variation (due to thermal subsystem aging); irreversible changes in units
Attitude disturbances	Antenna gain variation

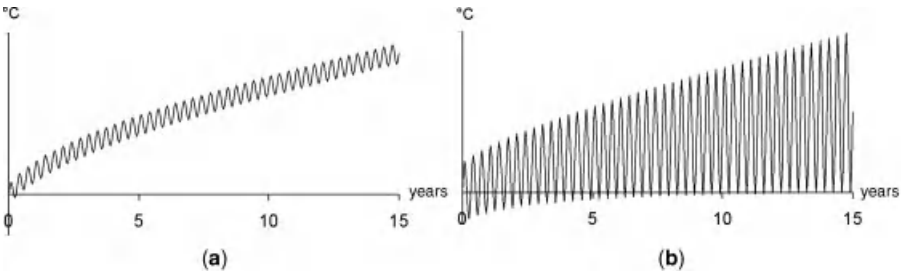


FIGURE 2.12 Examples (exaggerated) of unit temperature variation on a GEO: (a) unit on N or S panel when panels are thermally coupled and (b) unit on E or W panel when panels have independent heat pipe systems.

upward from **beginning of life (BOL)** to **end of life (EOL)**. Two examples of lifetime variation for a unit on a GEO satellite are shown in Figure 2.12, one example for a unit on the N or S panel and the other for a unit on the E or W panel. All panels are assumed to have heat pipes, and the N–S panels are thermally coupled. The yearly temperature variation is not shown; the diurnal variation is taken as the worst over the year. The N–S panels intrinsically provide a smaller diurnal temperature variation than independent E–W ones do.

2.3.1.1 On Active Units In general, active units have performance that varies with temperature. There are two groups of active units to consider: those that come before the preamp of the HPA subsystem and those that come after. The preamp divides the payload into two almost independent sides. The preamp must be able to accept a range of input power, part of which range accounts for temperature variation in units before it; additionally, it must be able to output a range of power to the HPA, part of which range accounts for temperature variation in the HPA.

Some payload units perform additional thermal control beyond what the bus's thermal-control subsystem does. Some units have heaters to increase their thermal stability. There is a type of TWTA that has its own radiator that extends out beyond the spacecraft edge; this provides relief to the thermal-control subsystem.

Some units correct their own performance over temperature based on uncorrected-performance curves measured in the factory.

2.3.1.2 On RF Lines As coaxial cable's temperature goes up, the insertion loss increases linearly; the insertion phase shift is a nonlinear function of temperature.

The behavior of waveguide over temperature depends on the waveguide material. Most materials expand linearly with increasing temperature, in which case the operating frequency range shifts down. Super Invar scarcely changes shape at all over a limited temperature range.

2.3.1.3 On Passive Filters The three filter technologies we will study are the waveguide filter, the dielectric-resonator filter, and the coaxial-cavity combline

TABLE 2.2 Summary of Payload Radiation Effects and Causes

Radiation Effect	Causes	Source
TID	Electrons, protons; X-rays from electrons	Trapped protons and electrons, solar protons; secondary radiation
SEE	Protons and heavier ions	Trapped protons, solar energetic particles, cosmic-ray particles

filter. The waveguide filter behaves as its waveguide material behaves. The other two are minimally temperature-sensitive.

2.3.2 Radiation Effects

The two types of radiation effects on current payload technology are "total ionizing dose" (TID) and "single-event effects" (SEE). The effects and their causes are listed in Table 2.2. Total ionizing dose is the total on-orbit accumulation of ionizations in semiconductor materials or associated insulators. Single-event effects are disruptions or permanent damage due to single incident particles (ECSS, 2008b). These effects are almost exclusively confined to digital electronics. The third type of radiation effect, displacement damage, is not found in today's payloads; it only applies to integrated circuits employing bipolar-junction transistors (ECSS, 2008b).

GaAs- and InP-based monolithic microwave integrated circuits (MMICs), commonly used for RF applications, are not significantly radiation-sensitive (Kayali, 1999). One supplier advises unit designers that their GaAs MMICs can take up to 10^7 to 10^8 rads total ionizing dose (TriQuint, 2010a). The transistors are high-electron-mobility transistor (HEMT), pseudo HEMT (pHEMT), or other field-effect transistors (FETs).

The only radiation-sensitive payload components are quartz crystal oscillators and integrated circuits (ICs) in complementary metal-oxide semiconductor (CMOS) technology, as shown in Table 2.3. Table 2.3 also shows which of the two radiation effects applies. A crystal changes due to total ionizing dose of radiation. In CMOS ICs, single-event effects include single-event upset (SEU), which is the nondestructive flipping of a bit, and single-event latch-up (SEL), which can cause part failure (ECSS, 2008b).

TABLE 2.3 Summary of Radiation Effects on Components

Component	Technology	Radiation Effects
Crystal oscillator	Oscillator crystal	TID
Integrated circuit	Bulk CMOS	TID and SEE
	Silicon on insulator (CMOS)	TID and SEE except for single-event latch-up

TABLE 2.4 Examples of CMOS IC Technology for Payload

IC Family	IC Function	Example of Usage
Digital	FPGA	On-board processing including packet processing and control, and DSP, and RAM (Xilinx, 2010)
	FPGA	(BAE, 2010)
	ASIC	(BAE, 2010)
	EEPROM	TCXO (Silanna, 2010)
	Microprocessor or microcontroller	TWTA EPC telemetry and control (Delporte et al., 1998)
Mixed-signal (i.e., analog/digital)		ADC, DAC (ECSS, 2008b)
RF/digital		Digital step attenuator and RF C-switch used in phased arrays (Robinette, 2009)
RF/analog/digital		PLL synthesizer (Peregrine, 2010)

Table 2.4 lists some examples of CMOS integrated-circuit technology which are in use today.

A payload unit may mitigate radiation by incorporating shielding. All components that are radiation-sensitive have been individually tested in radiation.

2.3.3 Aging Effects

Irreversible changes over life cause slow changes in unit performance even when the unit operating conditions remain the same. Aging is from accumulated operating time (ECSS, 2006).

Most RF GaAs-device failures occur in the FET channel, where failure is defined as reaching a 1 dB degradation in RF output power. Failure is from physical and chemical processes (TriQuint, 2010b).

What ages in a crystal oscillator is the crystal itself, its mounting, its electrodes, and so on (HP, 1997).

The lifetime drift of some payload-units' performance can be compensated through ground command. The supplier of any unit will have performed life test on several examples of that unit. Suppose that the supplier has observed a nonzero average drift per year. Then the future drift of a flight unit can be predicted but with uncertainty. Over the years, maybe the operation of the unit can be partially compensated through a commanded change in some parameter. However, if the drift observed by the supplier has a zero average, nothing can be done.

2.3.4 Antenna Gain Variation

Spacecraft attitude error causes antenna gain variation because the whole antenna pattern moves off. This is an especially difficult problem to analyze for a spot-beam payload.

REFERENCES

BAE Systems (2010). Radiation-hardened electronics. On www.baesystems.com/Products-Services. Accessed Nov. 26, 2010.

Briskman RD and Prevaux RJ (2001). S-DARS broadcast from inclined, elliptical orbits. *International Astronautical Congress;* Oct. 1–5; 503–518.

Cullimore BA (1992). Heat transfer system having a flexible deployable condenser tube. U.S. patent 5,117,901. June 2.

Daly EJ, Lemaire J, Heynderickx D, and Rodgers DJ (1996). Problems with models of the radiation belts. *IEEE Transactions on Nuclear Science;* 43; 403–415.

Delporte P, Fayt P, Gak M, and Pequet E (1998). EPC and TWTA for telecommunication satellites. *Proceedings of European Space Power Conference;* Sept. 21–25; 305–310.

Elbert BR (2004). *The Satellite Communication Applications Handbook*, 2nd ed. Boston: Artech House, Inc.

ECSS Std ECSS-Q-30-11A (2006). *Space Product Assurance, Derating—EEE Components*. The Netherlands: ESA Publications Division; Apr. 24.

European Cooperation for Space Standardization Std ECSS-E-10-04B draft (2008a). Space engineering, *Space Environment; Draft v0. 10*. The Netherlands: ESA Requirements and Standards Division; Mar. 13.

ECSS Std ECSS-E-10-12 draft (2008b). Space engineering, *Methods for the Calculation of Radiation Received and its Effects, and a Policy for Design Margins; Draft v0.20*. The Netherlands: ESA Requirements and Standards Division; Mar. 5.

Evans BG, editor (1999). *Satellite Communication Systems*, 3rd ed. London: The Institution of Electrical Engineers.

Hewlett Packard (1997). Fundamentals of quartz oscillators. Application note 200-2. On www.hpmemory.org/wb_pages/wall_b_page_01.htm. Accessed Nov. 24, 2010.

Hosick DK, inventor; Space Systems/Loral, Inc, assignee (2000). High power spacecraft with full utilization of all spacecraft surfaces. U.S. patent 6,073,887. June 13.

Jansson G, Intelsat General Corp. (2011). Use of commercial assets to host future communications systems. On www.intelsatgeneral.com/resources/white-papers. Accessed Nov. 21, 2011.

Jet Propulsion Laboratory (1999). Quicklook: Iridium. On msl.jpl.nasa.gov/QuickLooks/iridiumQL.html. Accessed Apr. 12, 2010.

Jondeau L, Flemin C, and Mena F, inventors; Astrium SAS, assignee (2010). Device for controlling the heat flows in a spacecraft and spacecraft equipped with such a device. U.S. patent application publication U.S. 2010/0001141 A1. Jan. 7.

Kayali S (1999). Reliability of compound semiconductor devices for space applications. *Microelectronics Reliability;* 39; 1723–1736. On trs-new.jpl.nasa.gov/dspace. Accessed Dec. 1, 2010.

Mansfield RL (2008). Predicting Iridium flares. *Poster presentation. American Astronomical Society, Division on Dynamical Astronomy Meeting;* Apr. 28–May 1.

Maral G and Bousquet M (2002). *Satellite Communications Systems*, 4th ed. Chichester, England: John Wiley & Sons, Ltd.

Martin DH, Anderson PR, and Bartamian L (2007). *Communications Satellites*, 5th ed. El Segundo, CA; Reston, VA: The Aerospace Press; American Institute of Aeronautics and Astronautics, Inc.

Morgan WL and Gordon GD (1989). *Communications Satellite Handbook*. New York: John Wiley & Sons.

O3b Networks Limited (2010). O3b Networks raises total funding of US $1.2 billion. Press release Nov. 29. On o3bnetworks.com/Media_Centre. Accessed Dec. 6, 2010.

Peregrine Semiconductor Corp. (2010). High-reliability products. Product sheet. May. On www.psemi.com/pdf/sell_sheet-psg/73-0019.pdf. Accessed Dec. 6, 2010.

Robinette D (2009). UltraCMOS RFICs ease the complexity of satellite designs. *Microwave Journal;* Aug.

Rodden JJ, Furumoto N, Fichter W, and Bruederle E, inventors; Globalstar L.P. and Daimler-Benz Aerospace AG, assignees (1998). Dynamic bias for orbital yaw steering. U.S. patent 5,791,598. Aug. 11.

Selding PB (2007). Globalstar says service sustainable until new satellites arrive in 2009. *Space News;* Feb. 12.

Silanna Semiconductor (2010). CMOS-on-sapphire technology delivers new level of multi-GHz RF integration. White paper. On www.silanna.com/index.php?page=sos-book. Accessed Nov. 26, 2010.

TriQuint Semiconductor, Inc (2010a). Gallium arsenide products, designers' information. On www.triquint.com/prodserv/tech_info/docs/mmw_appnotes/designer_a.pdf. Accessed Dec. 7, 2010.

TriQuint Semiconductor, Inc (2010b). Micro-/millimeter wave reliability overview. On www.triquint.com/shared/pubs/processes/Micro_Millimeter_Wave_Reliability_Overview.pdf. Accessed Dec. 7, 2010.

Wade M (2010). SDS-3. On www.astronautix.com. Accessed Nov. 30, 2010.

Watts KP, inventor; Hughes Electronics Corp., assignee (1998). Spacecraft radiator cooling system. U.S. patent 5,806,803. Sep. 15.

Wertz JR and Larson WJ, editors (1999). *Space Mission Analysis and Design*, 3rd ed. Hawthorne, CA; New York: Microcosm Press; Springer.

Wikipedia (2010a). Cosmic ray. On en.wikipedia.org. Accessed Dec. 6, 2010.

Wikipedia (2010b). Solar wind. On en.wikipedia.org. Accessed Dec. 6, 2010.

Wikipedia (2011). Geomagnetic pole. Oct. 12. On en.wikipedia.org. Accessed Oct. 28, 2011.

Xilinx, Inc (2010). Xilinx launches first high-density, rad-hard reconfigurable FPGA for space application. News release. On press.xilinx.com. Accessed Nov. 26, 2010.

Yee EM, inventor; Space Systems/Loral, Inc, assignee (2002). Spacecraft multiple loop heat pipe thermal system for internal equipment panel applications. U.S. patent 6,478,258 B1. Nov. 12.

Yee EM, inventor; Space Systems/Loral, Inc, assignee (2003). Spacecraft multi-directional loop heat pipe thermal systems. U.S. patent 6,591,899 B1. Jul. 15.

Yee EM, Space Systems/Loral (2011). Private communication. Nov. 7.

CHAPTER 3

ANTENNA

3.1 INTRODUCTION

This chapter is on the payload unit which is the most eye-catching—the antenna. Along with the solar panels, the antennas make a satellite look like a satellite. Antennas easily excite enthusiasm in some people: most antennas are big, many of them move, and antenna deployment videos are exciting, but what is more fun is to watch an antenna engineer demonstrate a two-axis deployment with his arm.

This chapter applies, as does the rest of the book, to communications satellites on orbit today or about to be launched. Also included are antennas developed by satellite manufacturers, because if a satellite manufacturer has spent the time and money to develop an engineering model, the probability is high that the manufacturer flies it within a few short years of reporting on it.

By far most payload antennas are for communications links between satellite and ground, but a few are for satellite-to-satellite crosslinks, of which we will mention three in the course of the chapter.

Excellent line drawings of all communication satellites up to 2007, which clearly show the antennas, can be found in Martin et al. (2007), along with brief but comprehensive descriptions of the spacecraft.

Here is a brief explanation of antennas as background for the detailed description that the rest of the chapter provides. A **reflector** is commonly called a "dish." A reflector antenna may have one or two reflectors. If it has two, the larger one is the **main reflector** and the smaller one is the **subreflector**. The (main) reflector is the one that directly receives and/or transmits power over communications distances. A reflector antenna always has a **feed**; on receive a feed collects the RF power

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

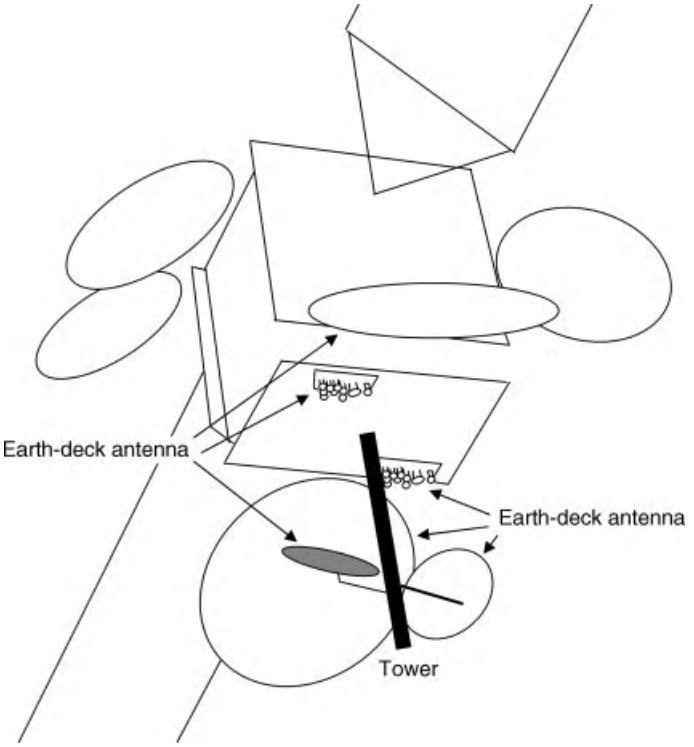


FIGURE 3.1 DirecTV-4S (after Boeing (2001)).

reflected into it by the reflector(s) and passes it to the front-end amplifier, and on transmit the feed sends out the RF power from a high-power amplifier to the reflector(s). A **horn** is a piece of waveguide, flared at the open end, that is most often used as a feed but is sometimes used as an antenna in its own right. An **antenna array** consists of many identical radiating elements in a grid pattern, and it forms beam(s) by combining elements in proper phase or timing relationships. Phase relationships are more commonly used, in which case the antenna array is a **phased array**. An antenna array is most often used to form the feeds for a reflector-based multibeam antenna but sometimes it is used as an antenna in its own right, when it is called an **array antenna**.

Figure 3.1 is a drawing of a more or less typical communications satellite, DirecTV-4S. The "S" in its name stands for "spot beam," and this was the first spot-beam satellite in the DirecTV fleet. The spacecraft was launched to GEO in 2001. It has five main reflectors. Two are attached to the earth deck, which faces down in the picture. Three other reflectors are on the two sides of the spacecraft that do not have solar panels, namely the east and west. All the dishes had to be deployed on orbit. The two earth-deck antennas are marked; each has a main reflector, a subreflector, and a feed cluster. The subreflectors are mounted on the **tower**. The antennas that

hang off the east and west panels do not have subreflectors, and their feeds are not visible.

The rest of the chapter is organized as follows:

- *Section 3.2:* Concepts that apply to any satellite communications antenna
- *Section 3.3:* Reflector antennas that create a single beam
- *Section 3.4:* Horn
- *Section 3.5:* Antenna array
- *Section 3.6:* Reflector-based multibeam antenna
- *Section 3.7:* Autotrack, by which an antenna autonomously maintains its pointing
- *Appendix 3.A*
  - *Section 3.A.1*: Meaning of dB
  - *Section 3.A.2*: Aperture antenna pattern
  - *Section 3.A.3*: Antenna-array antenna pattern.

3.2 GENERAL ANTENNA CONCEPTS

3.2.1 Beams

There are, roughly speaking, five sizes of antenna-beam coverage, defined the same for transmit and receive:

- **Global** or **earth-coverage** of the visible earth
- **Hemispherical**, which covers about half the visible earth
- **Zonal**, which covers approximately a continent
- **Regional** (also, **linguistic-region** in Europe), which covers an area about the size of Spain
- **Spot**, which covers an area about the size of metropolitan New York City. The term is sometimes used to mean instead a regional or zonal beam in comparison to a larger beam on the same satellite.

If the coverage area is not on the earth, the beam is named for the same solid-angular size as if it were on the earth. The Intelsat satellite fleet alone has all these sizes of beam (Schennum et al., 1999; Intelsat, 2008).

Especially at Ku-band and higher frequencies, the antenna may be designed to compensate long-term average attenuation due to rain. The compensation is done differently for different beam sizes: for global and hemispherical, probably no compensation; for zonal and regional, compensation via the beam providing higher gain toward some areas than others; and for spot, compensation via some beams having higher gain than others.

Transmit beams have power constraints. Toward their coverage areas, they must provide enough power but not too much or they violate **International**

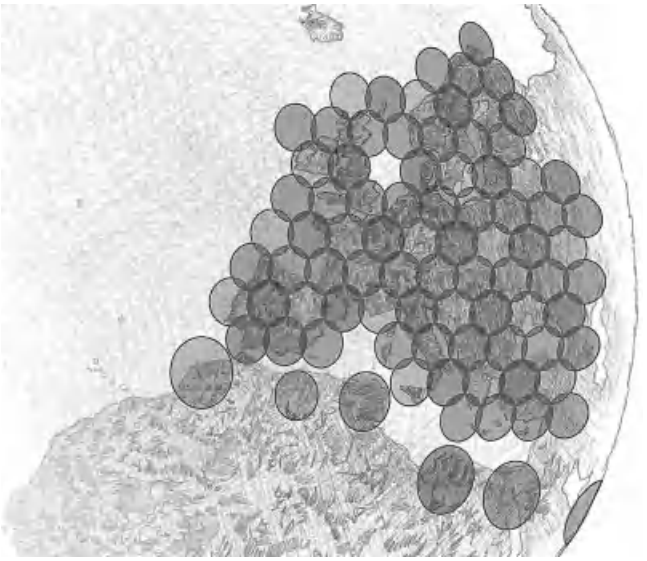


FIGURE 3.2 Spot-beam coverage of Eutelsat's Ka-Sat (KA-SAT drawing courtesy of Eutelsat).

Telecommunications Union (**ITU**) rules. Outside their coverage areas, in some cases beams must have very low power levels so as not to interfere into other countries' telecommunications.

Multiple spot beams are sometimes laid down on the earth like a tile floor, uniformly covering a large area in the way that Eutelsat's Ka-Sat covers Europe, as shown in Figure 3.2. This satellite was launched at the end of 2010. The spot beams form a hexagonal pattern, which flares out at the edges where the earth curves away from the satellite. The beam size is sometimes defined as the diameter of the beam at the triple-beam cross-over point; a typical value for the gain here is 4 dB below the gain at beam peak (see Appendix 3.A.1 for explanation of dB). The center-to-center spacing of adjacent beams is then 0.87 of the beam diameter, and a typical gain at the intersection point is 3 dB down from gain at beam peak (Rao, 1999). For some satellites, such as the Ciel-2 satellite, the spot beams do not fully cover CONUS but do cover the highly populated areas. Ciel-2 was launched in 2008. Each of its beams covers an area where people watch the same local TV stations, for example, sports shows.

Satellite beams may be differentiated from each other by three parameters: coverage area (also called "spatial separation"), antenna polarization (Section 3.2.3.3), and frequency bands carried. Overlaps in at most two of these parameters are allowed. The minimum feasible spatial separation of two beams with the same frequency band and polarization has long been regarded as two beams separated by another one (Cho et al., 1985; SGM, 2000). The minimum feasible frequency separation of two beams with the same coverage and polarization is 10% of the channel bandwidth between channel-band edges, called the "guard band" (illustrated in

Section 4.5.3.4). Polarizations must be orthogonal for two beams of the same coverage and frequency band. The interference created by the always imperfect differentiation in the three parameters is dealt with in Section 11.5. In the case of densely packed spot beams, the number of frequency bands used times the number of polarizations used in covering the area has been called the number of "colors"; for example, two bands and two polarizations is a "four-color" scheme (Maufroid et al., 2004). The number of times each color is used in the total coverage is the **frequency reuse factor**.

3.2.2 Aperture

All the kinds of payload antennas, namely reflector antennas, horn antennas, and planar arrays, belong to the family of aperture antennas (Chang, 1989). An aperture antenna is one for which a plane can be defined on which the tangential electric or magnetic field strength distribution is known or can be well estimated; the field is significant over only a finite area in this plane, and this finite area is the **aperture** (Chang, 1989). For a reflector antenna with a main reflector which is the surface of revolution of a two-dimensional curve, the aperture is the flat circular area that would close off the reflector surface; the aperture is perpendicular to the curve's axis of revolution. When the main reflector's surface is just part of such a symmetric surface, the aperture is the projection of the reflector surface onto the same flat circular area. For a horn, it is the radiating opening. For a planar array, it is the part of the plane containing the array.

The aperture's tangential electric field determines the electric field everywhere in the half-space in front of the aperture, and the aperture's tangential magnetic field determines the magnetic field everywhere there (Chang, 1989). If we call the aperture plane the $z = 0$ plane, we can say that all points with $z > 0$ are in the half-space in front of the aperture. An implication is that there is no radiated electric field in the half-space behind the aperture. An actual aperture would in fact have a little bit of **spill-over**, where some radiation curls back around the aperture edges, but it is usually negligible by design.

The tangential electric field is an RF vector field so has the properties of direction, phase, and amplitude averaged over a cycle. The aperture is said to have **uniform illumination** if phase and average amplitude are constant across the aperture (Collin, 1985). It is said to have **tapered illumination** if the amplitude progressively rolls off approaching the edges.

3.2.3 Antenna Pattern

An antenna pattern has two parts, the gain pattern and the polarization pattern. The antenna pattern is defined at every direction from the antenna.

The **field of view** (**FOV**) of an antenna can be defined in several ways (Gagliardi, 1978), but we define it as those directions over which the antenna pattern must satisfy minimum requirements. **Edge of coverage** (**EOC**) is the outline of the FOV on the earth.

Passive antennas have the **reciprocity property**, whereby they have the same pattern on receive and transmit (Collin, 1985). In fact, passive receive antennas are tested as transmit antennas at the receive frequency. Passive antennas do not have amplifiers incorporated into them, as some antenna arrays do, at each radiating element.

3.2.3.1 Gain, EIRP, and G/T$_s$ The purpose of a payload-transmitting antenna is to shape the radiation of the RF power into the antenna terminal so it goes preferentially in some directions and much less in the other directions. Similarly, a payload receiving antenna shapes the receiving sensitivity of the payload to maximize it in the directions desired and suppresses it in the other directions. This shaping is characterized by the antenna **directivity function** $D(\theta,\phi)$ defined on all polar angles θ and **azimuths** ϕ. The polar angle is more commonly called the **off-boresight angle**. The angles of a direction in the spherical coordinate system are defined in Figure 3.3, where the antenna aperture lies in the $z = 0$ plane and ϕ is measured in the counter-clockwise direction about the z-axis.

It is easiest to understand what directivity is if we first think about the **ideal** (theoretical, nonexistent) **isotropic antenna**. It is a point source and has unity directivity in every direction. It does not shape the radiation but sends equal amounts in all directions. Its integral over all directions is given by the following:

$$\int_0^{\pi} d\theta \int_0^{2\pi} D(\theta,\phi)\sin\theta d\phi = \int_0^{\pi} d\theta \int_0^{2\pi} \sin\theta d\phi = 4\pi \quad \text{for isotropic antenna}$$

For the directivity function of any other antenna, the value of the integral is the same:

$$\int_0^{\pi} d\theta \int_0^{2\pi} D(\theta,\phi)\sin\theta d\phi = 4\pi \quad \text{for any antenna}$$

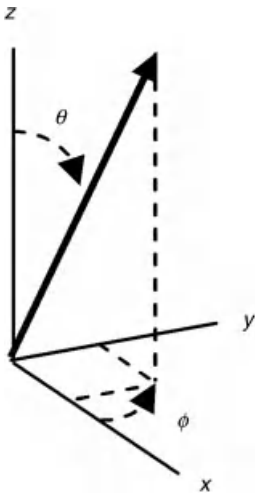


FIGURE 3.3 Angles of spherical coordinate system.

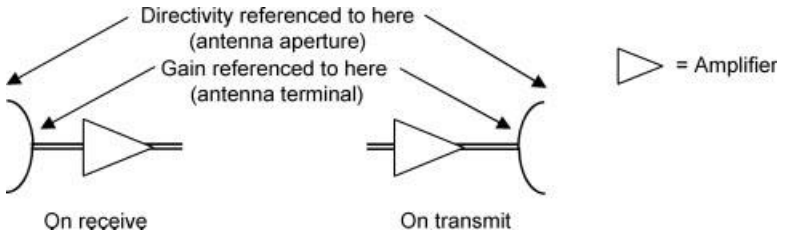


FIGURE 3.4 Reference points of directivity and gain.

which means that if there are directions in which the antenna's directivity is large, there must be compensating directions in which the directivity is small, less than unity. For aperture antennas, neglecting any spillover (Section 3.2.2), D is zero for all $z < 0$ or $\theta > \pi$.

While the antenna is accomplishing its purpose of shaping the radiation, it must also cause minimal loss. Directivity is referenced to the total power radiated out of the antenna, but the more useful **gain** is referenced to the power into the antenna terminal. It is more useful because it characterizes the antenna as a whole, that is, as a unit that connects at its terminal to another unit, as shown in Figure 3.4. Gain is proportional to directivity by a factor η that accounts for antenna losses, called the **antenna efficiency**, which will be discussed in more detail in Section 3.2.7 but is simply represented here

$$P_{\text{radiated}} = \eta P_{\text{terminal}} \qquad \text{for transmitting antenna}$$

$$P_{\text{terminal}} = \eta P_{\text{received at aperture}} \qquad \text{for receiving antenna}$$

Normally gain is spoken of not in terms of a numerical ratio but in decibel (dB). All payload antennas are high-gain, since even the payload antenna with the lowest gain, the segment antennas on Globalstar, have a gain of about 20 dB. The lowest-gain payload antennas are on LEOs.

EIRP, **equivalent isotropically radiated power**, is the most important parameter of the majority of payloads, since most payloads broadcast. It comes from a combination of the power into the transmitting antenna's terminal and the antenna gain. EIRP is defined on all directions. EIRP in a given direction is the total power that the antenna would radiate if it could have the gain in all directions that it has in the one direction:

$$\text{EIRP}(\theta, \phi) = G(\theta, \phi) P_{\text{terminal}}$$

G/T_s, the **figure of merit** of the payload in its function as a receive terminal, is the most important parameter for satellite uplinks. It comes from a combination of the receive antenna gain and the system noise temperature, T_s (Section 7.12). The parameter sets the uplink signal-to-noise ratio (SNR).

3.2.3.2 Far Field and Near Field The antenna pattern exists in the **far field** of the antenna. There, at communications distances, the antenna's radiation is

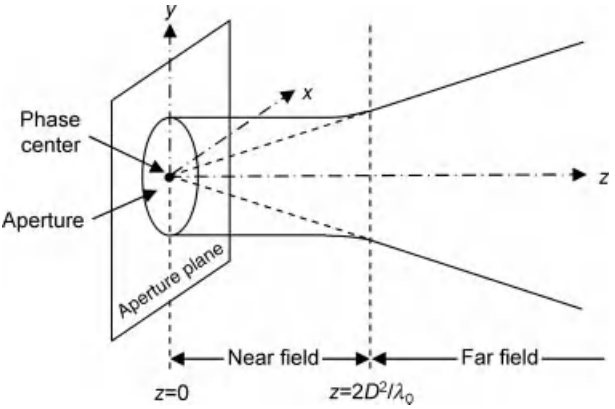


FIGURE 3.5 Near- and far-field regions for high-gain aperture antenna (after Chang (1989)).

effectively a planar wavefront with the **transverse electromagnetic** (**TEM**) propagation mode. In TEM, the electric field vector ***E*** and its orthogonal magnetic field vector ***H*** lie in the wavefront plane, that is, are perpendicular to the direction of propagation (Collin, 1985), and the fields' amplitudes are related by a constant (Ramo et al., 1984). The far field for a high-gain aperture antenna is generally said to start at a distance of $2D^2/\lambda_0$ from the aperture, where D is an effective diameter of the antenna aperture and λ_0 is the wavelength of the center frequency being transmitted or received (Chang, 1989). The wavefronts, which have constant phase by definition, seem to radiate from one spot, the antenna's **phase center**. This is illustrated in Figure 3.5. In this example, the phase center lies in the aperture but this is not always the case, for example, for horns (Section 3.4.3).

The antenna pattern of an aperture antenna can be derived from the two-dimensional Fourier transform of the tangential electric field in the aperture (Collin, 1985). At small off-boresight angles, the pattern very nearly equals the transform. The larger the aperture the narrower the beam, so the higher the gain at peak of beam (Appendix 3.A.2). The antenna pattern of an array antenna is approximately given by the product of an individual array element's pattern and the **array factor**, which is related to the array's layout. This is discussed in Appendix 3.A.3.

From Figure 3.5, we see that besides the far field there is the **near-field** or **Fresnel region**. Here, the beam diameter increases slowly and the cross-sectional amplitude and phase distributions change little with distance z from the aperture (Chang, 1989).

Often when an antenna is tested in its far field, measurements are only taken in directions that lie in two orthogonal planes (see Appendix 7.A.1 for antenna testing). The planes are the **principal planes**, specifically the ***E*-plane** and the ***H*-plane**. Figure 3.6 illustrates their definition, assuming that in the far field in a plane perpendicular to the peak-of-beam direction, the ***E*-** and ***H*-**field vectors point as shown. Each plane contains the vector to the peak of beam. The pattern is said to be taken in an ***E*-**plane or ***H*-**plane **cut**, that is, the set of directions from the phase center that

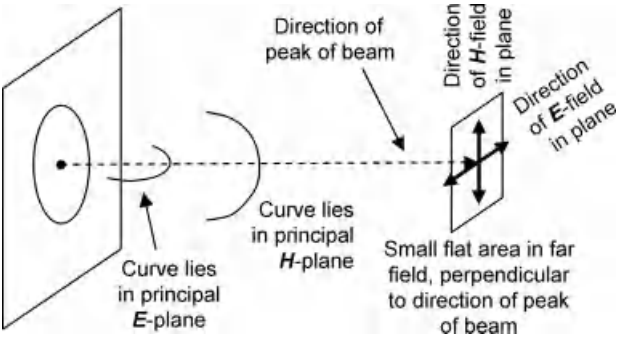


FIGURE 3.6 Principal planes for antenna measurements.

lie in the corresponding plane. Other cuts, in other convenient planes, can be defined.

3.2.3.3 Gain Pattern A gain pattern for communications has a mainlobe and sidelobes separated by nulls, as the example in Figure 3.7 shows. (This example is unrealistically bad but illustrates the definitions well.) The mainlobe is much higher than the sidelobes. There are approximate expressions for the null-to-null width of the mainlobe (Collin, 1985). For a rectangular horn whose aperture in the wide direction has dimension D, the null-to-null beamwidth in a cut in the wide direction is as follows:

$$\text{Mainlobe beamwidth} \approx \frac{\lambda_0}{D/2} \quad \text{rad}$$

For a circular aperture of diameter D, the expression is as follows:

$$\text{Mainlobe beamwidth} \approx \frac{3.8}{\pi}\frac{\lambda_0}{D/2} \quad \text{rad}$$

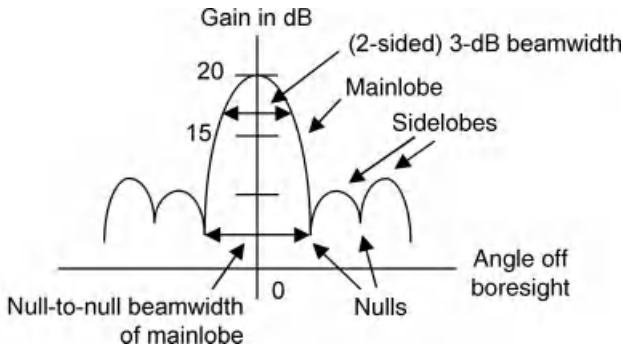


FIGURE 3.7 Definition of some antenna-gain pattern terms.

Thus, the larger the aperture is in units of wavelength, the narrower the mainlobe. The (two-sided) half-power (i.e., 3 dB down from peak of beam) beamwidth of a circular aperture with uniform illumination is as follows (Evans, 1999):

$$\text{Half-power beamwidth} \approx \frac{\lambda_0}{D} \quad \text{rad}$$

(Compare this with the half-power width of the transform of the rectangular pulse in Section 10.3.4.2.)

3.2.3.4 Polarization Polarization of the far-field electric field is the other property of an antenna pattern, besides gain.

When an antenna radiates, at any given point in space the electric field vector moves repetitively with the frequency of the radiation. The general shape that the vector traces is an ellipse, and the ellipse possibilities range from a circle to a line. Looking in the direction of propagation, if the direction of rotation is clockwise the radiation is said to be **right-hand polarized** (**RHP**), and if counter-clockwise it is said to be **left-hand polarized** (**LHP**) (Collin, 1985). If the ellipse is in fact a circle, the radiation is either **right-hand circularly polarized** (**RHCP**) or **left-hand circularly polarized** (**LHCP**). If the ellipse is merely a line that gets traced back and forth, the radiation is **linearly polarized** (**LP**). The most common linear polarizations are **horizontal (H)** and **vertical (V)**, where horizontal has the $\boldsymbol{E}$-field parallel to the ground (on the earth) and vertical is orthogonal to horizontal and to the direction of propagation. The **axial ratio** r of an ellipse is defined as the ratio of the minor axis of the ellipse to its major axis, which are illustrated in Figure 3.8. Thus, $r = 1$ is a circle and $r = 0$ is a line. Axial ratio in dB is $-20 \log_{10} r$.

At C-band and lower frequencies, circular polarization is most often used for satellite communications, but at Ku-band and higher frequencies both circular and linear are used (Hoffmeister, 2010). The reason is Faraday rotation (Section 11.3.4.1).

For the receiving antenna to be able to receive all the radiation incident on it from a transmitting antenna, the radiation and the receiving antenna must have the same polarization (Chang, 1989). If the two polarizations are orthogonal, no radiation will be received. Examples of pairs of orthogonal polarizations are (1) RHCP and LHCP and (2) vertical and horizontal. If a CP antenna receives LP, it receives only half of the power, and vice versa.

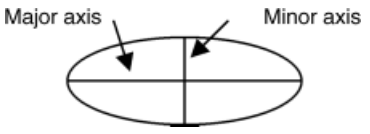


FIGURE 3.8 Ellipse axes.

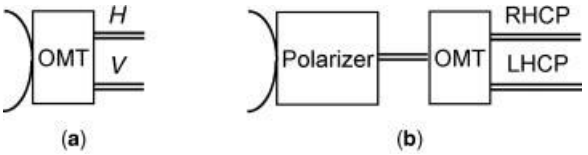


FIGURE 3.9 (a) Orthomode transducer (OMT) providing dual LP; (b) polarizer and OMT providing dual CP.

3.2.4 Orthomode Transducer and Polarizer

An **orthomode transducer** (**OMT**) is a device that allows for receiving or transmitting dual LP instead of just a single sense of LP. On receive, it separates out and converts two orthogonal TE_{11} modes in one circular waveguide to two orthogonal TE_{10} modes in two rectangular waveguides (Schennum et al., 1999). **TE** is **transverse electric** and means that the electric field is perpendicular to the direction of propagation and that the magnetic field is not (for discussion on propagation modes, see Section 4.3.2.5). On transmit the OMT does the opposite of what it does on receive. On each LP sense it can either receive or transmit, independently of what it is doing on the other LP sense, as shown in Figure 3.9a. In Figure 3.9, *H* and *V* are not necessarily meant here to be horizontal and vertical, just an orthogonal pair of linear polarizations. The OMT's important parameter is its port-to-port isolation.

A **polarizer** is a device for receiving a CP signal and converting it into a LP signal. On transmit it does the opposite. It can also act on dual polarizations. On each CP sense it can either receive or transmit, independently of what it is doing on the other CP sense, as shown in Figure 3.9b. The important parameter of a polarizer is its cross-polarization isolation.

On receive the polarizer changes a single-sense CP electric field into a single sense of LP by separating out orthogonal linear components and delaying one, as illustrated in Figures 3.10 and 3.11, respectively. The component drawn horizontally is delayed by a quarter cycle between the two figures.

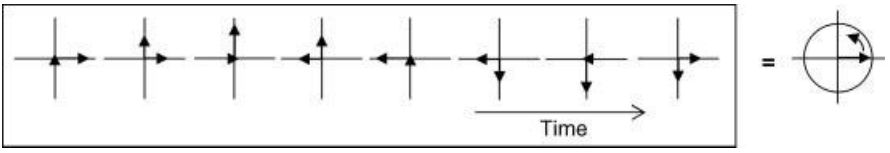


FIGURE 3.10 Resolution of one CP sense into two senses of LP.

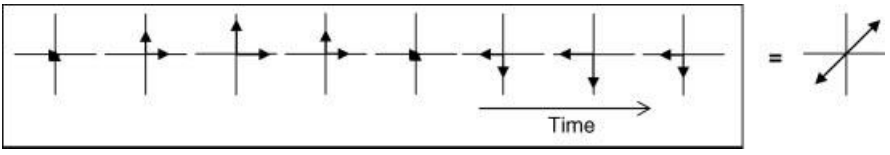


FIGURE 3.11 One LP sense delayed; combination of two senses resulting in one LP sense at 45°.

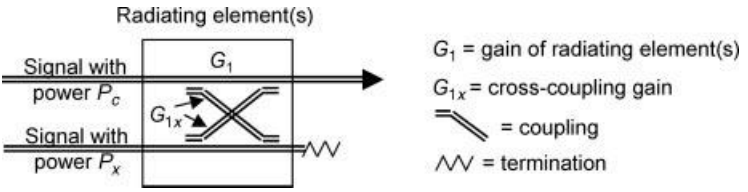


FIGURE 3.12 Cross-polarization interference shown as cross-coupling, for case where only one polarization is intended to be received.

3.2.5 Cross-Polarization Induced by Antenna

In creating radiation of a desired polarization, a small amount of orthogonally polarized radiation is also inevitably created. The radiation with the desired polarization is said to be **copolarized** or simply **co-pol**, while the undesired radiation is said to be **cross-polarized** or **cross-pol**. The only time cross-pol is a concern is when it causes interference, so cross-pol and **cross-pol interference** are practically synonymous. If the cross-pol interference and signal it is interfering with are unrelated, the interference is modeled as additive noise.

Cross-pol has several sources on the satellite. On receive the first source is the radiating element(s). For a reflector antenna this would be the reflector, subreflector if any, and the feed. For a horn antenna, it is the horn itself. For an antenna array, this would be all the radiating elements that are combined for the beam. For any antenna, cross-pol could also arise from reflections off struts or anything else, but with a good bus and payload design this is probably negligible. If the antenna is receiving dual CP then the next cross-pol source is the polarizer. If the antenna is receiving dual pol, the next and last source is the OMT. Figure 3.12 shows the situation for reception of one polarization and Figure 3.13 for reception of dual pol, where at each element the creation of cross-pol is drawn as cross-coupling. Cross-pol arises on transmit from the same sources.

Let us quantify the cross-pol interference in the case corresponding to Figure 3.13. The two arriving signals are assumed to be perfectly orthogonal (the imperfect case is dealt within Section 11.5.4). In the through paths both signals receive the same gain. Within each of the three elements a bit of each signal is cross-coupled into the other signal. The cross-pol performance of an element is normally quoted as "relative gain" or is in dBc, which are the same. The relative gains for the three

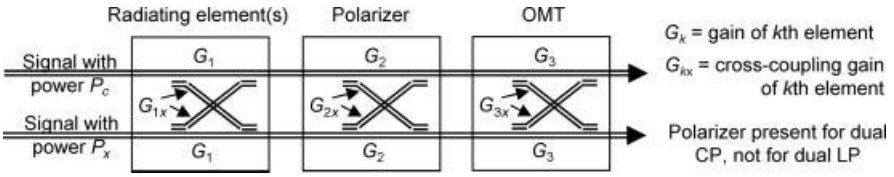


FIGURE 3.13 Cross-polarization interference on receive shown as cross-coupling.

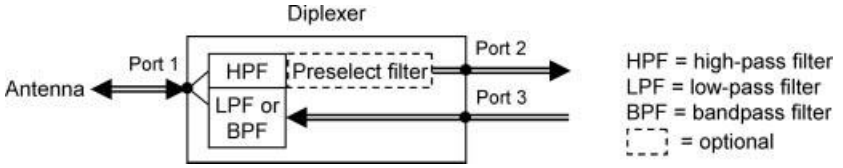


FIGURE 3.14 Diplexer for receiving and transmitting on same antenna, in one polarization.

elements are the gain ratios $\gamma_k = G_{kx}/G_k$ for $k = 1, 2, 3$, which are not in dB. We find that

$$\frac{P_c}{(\gamma_1 + \gamma_2 + \gamma_3)P_x} = \text{ratio of co-pol power to interference power from cross-pol signal}$$

$$\frac{P_x}{(\gamma_1 + \gamma_2 + \gamma_3)P_c} = \text{ratio of co-pol power to interference power from cross-pol signal, for other signal}$$

3.2.6 Diplexer

An antenna can both receive and transmit at the same time by means of a **diplexer**. The same antenna can both receive and transmit in one transponder or can receive in one transponder and transmit in another. The receive and transmit signals must be in different frequency bands.

A block diagram of a diplexer is shown in Figure 3.14. In the receive signal path, a piece of waveguide sized so that the transmit frequency is below cutoff, blocks the transmit signal and passes the receive signal (Schennum, 2011). This waveguide is effectively a high-pass filter (HPF). Because the transmit signal is very much stronger than the receive signal, the isolation from port 3 into port 2 must be very high. The payload's preselect filter, which is a bandpass filter (BPF), may follow the HPF in the diplexer (Kwok and Fiedziuszko, 1996) or it may be separate from the diplexer (Schennum et al., 1995). In the transmit signal path is a low-pass filter or a BPF. There can be two such diplexers per antenna, one for each polarization, as shown in Figure 3.15.

3.2.7 Losses and Antenna Temperature

The total power that an antenna radiates is less than the amount of power input to the antenna terminal because the antenna efficiency is less than 100%, that is,

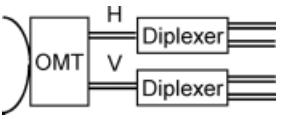


FIGURE 3.15 Two diplexers with OMT, for dual LP.

various losses occur in regard to the antenna. Similarly, for a receiving antenna the amount of power out of its terminal is less than the total amount of power reaching the aperture, because of similar losses. These losses can be identified as follows:

- Aperture or illumination inefficiency, resulting from deviation of aperture's tangential electric field from constant-amplitude and constant-phase across the aperture
- Blockage inefficiency, resulting from blockage of radiation by the feed, subreflector if any, and/or struts holding up the feed or subreflector
- Spillover inefficiency, resulting from some of the power radiated from the feed missing the reflector and/or subreflector. Also called "beam efficiency," from the viewpoint of the feed horn. There is an analog on receive, which would be some of the power collected by the reflector missing the subreflector or feed and some of the power collected by the subreflector missing the feed.
- Surface inefficiency, resulting from the reflector and/or subreflector surface deviating from what was designed, due to manufacturing and alignment inaccuracies
- Cross-polarization inefficiency, the loss of the power that is coupled into the orthogonal polarization
- Ohmic or resistive inefficiency, resulting from losses in the conducting surfaces
- Frequency dispersion, since the antenna is not quite as effective at other frequencies as at the frequency at which it was optimized
- Thermal distortion.

In addition, there are numbers that would be carried in the antenna loss budget but may or may not actually represent degradations in performance:

- Computer modeling error in design stage
- Measurement error.

Last, there are terms that only apply to an antenna array (see Section 3.5):

- Imperfect beam-forming
- Scan loss from frequency dispersion in scanned beams of a phased array.

There is also an impedance-mismatch loss between the antenna terminal and the connecting RF line, but this may be considered to be a payload-integration loss and not booked to the antenna. A mismatch of impedances at a connection causes part of a signal reaching the connection in either direction to be reflected (creating a return loss) (Sorrentino and Bianchi, 2010).

When an antenna receives, it contributes a noise temperature term to the total system noise temperature of the payload (Section 7.12.1). This **antenna temperature T_a** is not actually due to the antenna but to noise that the antenna receives

along with the uplink signal. The temperature is usually referenced to the antenna terminal so has been affected by the antenna's ohmic inefficiency (loss).

3.2.8 Reconfigurability

We will say that an antenna is **reconfigurable** if its antenna pattern(s) can be changed on orbit (this is not a universally accepted definition). Among reconfigurable antennas are reflector antennas with steerable beams and array antennas with flexible beam-forming. There are various reasons that reconfigurability may be desired, for example, on a GEO satellite, to shift beams to accommodate traffic variation, to keep aim at the Space Shuttle as it orbits, or to redefine the beams if the satellite is moved to another orbital slot; and for example, on a LEO satellite, to repoint inter-satellite links.

3.2.9 On-Orbit Environment and Its Mitigation

There are two aspects of the on-orbit environment that the antenna must be built to deal with: the thermal environment and the spacecraft pointing accuracy.

The antennas are on the outside of the spacecraft body so they see the full range of temperature, namely at different times they can be fully in the dark, partially lit by the sun, or fully lit by the sun at an oblique or boresight angle. One manufacturer always find that the worst thermal distortion, that is, the biggest rms surface error, is in the dark (cold) (Schennum, 2010b). Reflectors are made of a material that is thermally very stable, such as graphite. A gridded reflector, which actually consists of two surfaces (Section 3.3.1), may be encased by a protective cover which reduces temperature extremes and gradients, as on the Eutelsat II satellites of the early 1990s (Duret et al., 1989). The reflecting surfaces of reflectors may be coated with low-insertion-loss thermal paint, as on NASA's Advanced Communications Technology Satellite (ACTS) launched in 1993 (Regier, 1992). Thermal paint or coating makes good thermal contact with the antenna reflector, radiates heat well, and reflects light well (Wertz and Larson, 1999).

One phased array which is active on both receive and transmit, on the Iridium satellite (Section 3.5.3), has the active electronics on the back of the array panel. The heat from the high-power amplifiers (HPAs) is conducted via heat pipes to the front of the array, from which it radiates into space. The front of the array has a thermal coating. The active electronics must be heated during times of low communications traffic (Schuss et al., 1999).

The spacecraft's attitude control system may reduce the antenna pointing error to tenths of a degree, and antenna autotracking can reduce it to hundredths (Lepeltier et al., 2007). A separate onboard tracking antenna can also reduce it to hundredths (Amyotte et al., 2006). The antenna patterns must be adjusted to account for the pointing error by enlarging the beams and, for multibeam antennas, by ensuring that interference from other enlarged beams is at an acceptable level. The customer and manufacturer must agree on exactly what measure of pointing error is to be used, for example, 2-sigma.

3.3 SINGLE-BEAM REFLECTOR ANTENNA

3.3.1 Reflector Antenna Concepts

A reflector antenna has at least a main reflector and a feed. By far the most common form of main reflector is a paraboloid or section of a paraboloid (Chang, 1989). The general idea of a reflector antenna is that the feed either is or appears to be at the **focus** of the paraboloid, so that upon reflection the rays emerge parallel and in phase in planes perpendicular to the paraboloid's axis, as shown in Figure 3.16. A partial paraboloid also has this property.

If the antenna has only a main reflector and a feed, it is a **single-reflector antenna**, for which the feed is at the focal point. If additionally it has a subreflector, it is a **dual-reflector antenna**, and the feed is no longer actually at the focus but just virtually. In either case the phase center (Section 3.2.3.2) of the feed must be known so that the antenna geometry can be set up right. A reflector antenna is inherently wideband if the phase center of the feed or virtual feed is exactly at the focus (Chang, 1989).

If only an asymmetric part of the paraboloid is there, the antenna is **offset-fed** because the feed will not be set to point down the paraboloid's axis, but it will still be at the focus.

One of the defining characteristics of a reflector antenna is ***f*/*D***, the focal length f in units of the paraboloid diameter D. When the paraboloid is only partial, its f and D are still defined as if the paraboloid were all there. (However, gain is a function of the actual aperture diameter.) There does not seem to be a standard definition of what is a short f/D and what is a long f/D—one definition uses the dividing line value of 1 (Howard, 1983) and another uses 0.7 (Legay et al., 2000).

The illumination of the aperture is usually uniform because this yields maximum gain from the antenna (Chang, 1989). The **primary radiation pattern**, that is, that of the feed, has to be shaped to provide the same electrical field strength at the

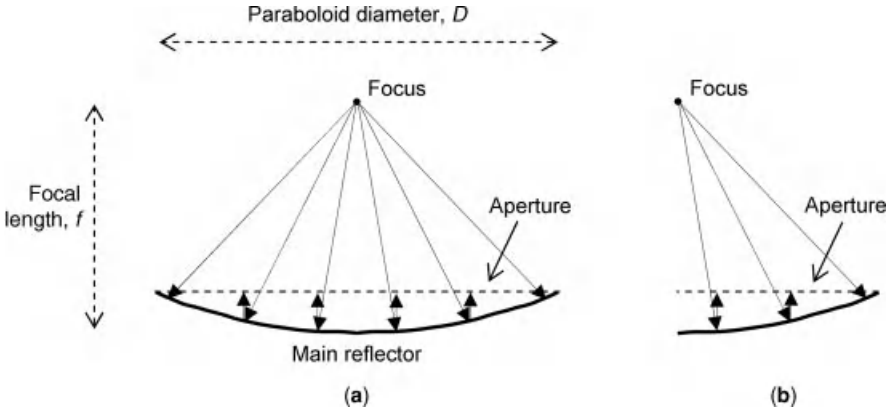


FIGURE 3.16 Basic property of paraboloid—all rays in phase at aperture: (a) full paraboloid and (b) partial paraboloid.

reflector edges as in the middle, to compensate the extra space loss to the edges (Chang, 1989). The greater *f/D* is, the less shaping of the primary radiation pattern is required (Collin, 1985). The reflector's pattern is the **secondary radiation pattern**.

The feed and any subreflector are in the near field of the reflector, where the radiation from or to the reflector is approximately straight out from the reflector aperture, as shown in Figure 3.5. From the point of view of the feed, the reflector is typically in its far field (Rudge et al., 1982), but in a dual-reflector antenna the subreflector may not be in the feed's far field (Albertsen and Pontoppidan, 1984; Rahmat-Samii and Imbriale, 1998).

The sense of circular polarization is reversed every time the signal is reflected, so a RHCP single-reflector antenna requires a LHCP feed.

The only type of feed in use today for single-beam reflector antennas is the single horn, which is discussed in Section 3.4. The ability to appropriately **shape** the surface of the main reflector and any subreflector has grown to the point where a phased array is no longer necessary to obtain a well-shaped beam contour (see Sections 3.3.2 and 3.3.3 for discussions on surface shaping).

A main-reflector design that creates very low cross-pol for dual LP separates the main reflector into two surfaces, one nested inside the other and turned at a slight angle so that the two feeds can be in slightly different places. The geometry is illustrated in Figure 3.17 for the offset-fed case. The front surface is **gridded**, that is, it is an array of conductors parallel to one linear electric field. Radiation with that polarization is reflected but the orthogonal LP sense passes through and is reflected on the back surface. This was thought of in the 1970s (Rudge et al., 1982). It was used in the antennas of the Eutelsat II satellites (Maral and Bousquet, 2002), launched from 1990 to 1995, and in the Ku-band spot beam antenna S2 on the Intelsat VIIA satellites (Schennum et al., 1995), launched in 1995 and 1996. The main reflector is **dual-gridded** if the back reflector is also gridded.

A reflector antenna can be made to have either one, a small number, or a large number of beams. When it creates one or a small number, the beams are most likely zonal beams, for example, a continent or a large part of a continent. When it creates a large number of beams, the beams are most likely spot beams. The problems of designing a reflector antenna with multiple beams are those of designing a single-beam antenna plus new concerns. In Section 3.4, we deal with the single-beam reflector antenna, and in Section 3.5 with the reflector-based multibeam antenna.

FIGURE 3.17 Gridded offset-fed main-reflector geometry.

3.3.2 Single-Reflector Single-Beam Antenna

The geometry of the **center-fed** paraboloidal reflector is shown in Figure 3.18a. The feed, which is at the focal point of the paraboloid, must radiate energy in all directions within the angle θ_{sub}, which the reflector subtends from the focus. The angle θ_{sub} is a function of f/D.

The center-fed reflector does not have especially low levels of cross-pol for LP, but it theoretically has no cross-pol with a CP feed (Chang, 1989). The center-fed reflector has the deficiency that part of the reflector is blocked by the feed, the feed's supporting struts, and the RF line to the feed. The scattered radiation causes not just loss of gain but also higher sidelobes (Chang, 1989).

A solution with no blockage is the offset-fed paraboloidal reflector, with the geometry shown in Figure 3.18b. The feed is still at the focus but is now pointed in an offset direction toward approximately the middle of the reflector surface (Chang, 1989). A larger f/D is feasible than with center-fed, which means larger-aperture feeds (to create a narrower beam), which can be made to have a better-shaped radiation pattern and lower cross-pol. Another advantage of the offset configuration is that the reflector does not reflect radiation back into the feed or feeds (Rudge et al., 1982). A disadvantage is that the cross-pol is increased for LP (Chang, 1989), but the feed can compensate this (Section 3.4). The offset-fed reflector theoretically has no cross-pol with a CP feed, but the beam from the main reflector **squints**, that is, its boresight direction is off, in a direction perpendicular to the feed's pointing offset (Chang, 1989).

For both center-fed and offset-fed antennas, the reflector can be shaped to improve the beam's contour. The outline of the reflector can be changed and/or the surface can be shaped. A tradeoff study was conducted on how to obtain a 12.5 GHz CONUS beam from GEO with 1.5 dB variation of gain across CONUS to compensate rain attenuation, with an offset paraboloidal single-reflector antenna. It was found that surface shaping and a single feed could accomplish this. The surface deviated from a paraboloid by 1.4 in. (about 1.5 wavelengths) at most. Without surface shaping a 56-element phased-array feed would have been necessary (Ramanujam et al., 1993).

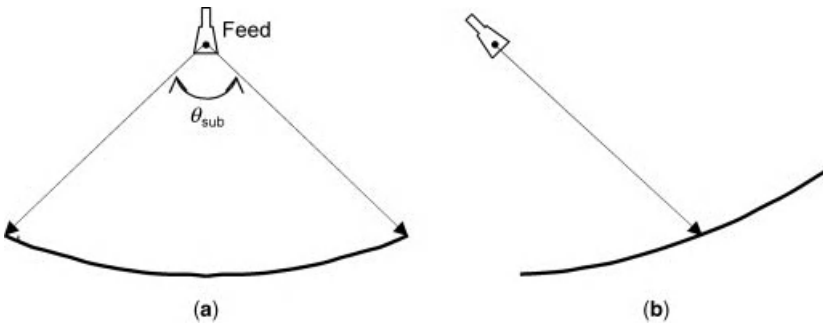


FIGURE 3.18 Reflector and feed of paraboloidal antennas: (a) center-fed and (b) offset-fed.

The three ways to make a single-reflector antenna reconfigurable are to gimbal the whole antenna, just the reflector, or just the feed. The last two are mechanically simpler than the first. The second way incurs scan loss due to the feed no longer being at the reflector focal point (Choung, 1996).

3.3.3 Dual-Reflector Single-Beam Antenna

3.3.3.1 Center-Fed The dual-reflector antenna has a subreflector in addition to the main reflector, and the feed is not at the paraboloid's focus. One type of dual-reflector antenna is the **Cassegrain** design, which has long been used for optical telescopes—the design was first published in 1672. The hyperboloidal subreflector has two foci, one inside the hyperboloid's curve and one outside, as shown in Figure 3.19a. The geometrical arrangement of the Cassegrain is that the hyperboloid's inside focus is set at the main reflector's focus and the outside focus is set at the feed's phase center, as shown in Figure 3.19b. The hyperboloid can be designed so that the feed is slightly behind the main reflector and therefore causes no blockage. The Cassegrain has other configurations than the classical one shown in the figure (Chang, 1989).

The other type of dual-reflector antenna used is the **Gregorian** design, which has also long been used in optical telescopes, being first published in 1663. Its subreflector is the small end of an ellipsoid. The two foci of the ellipsoid are set with one at the paraboloid's focus and the other at the feed's phase center. Figure 3.20 shows a Gregorian antenna with a reflector of the same diameter as the Cassegrain's in Figure 3.19, namely with the same blockage of the main reflector. Compared to the Cassegrain, the Gregorian's subreflector is farther away from the main reflector so requires longer supporting struts and the feed has a narrower beam so must be larger. To first order the performance of the Gregorian is the same as that of a Cassegrain with the same f/D and D. The Gregorian has other configurations besides the classical one shown in the figure (Chang, 1989).

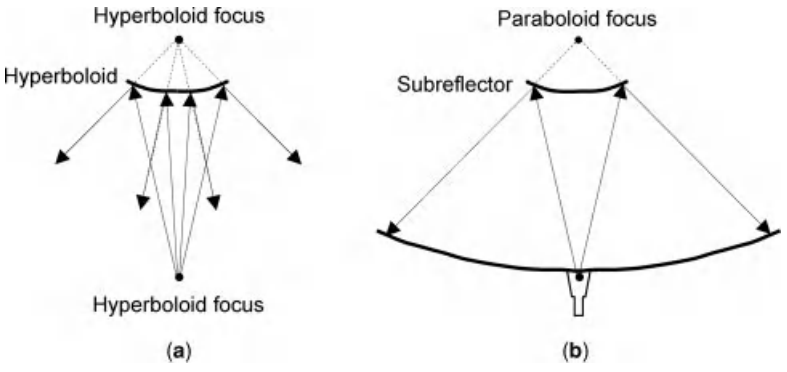


FIGURE 3.19 (a) Basic property of hyperboloid—all reflected rays from one focus seem to come from other focus and (b) center-fed Cassegrain antenna.

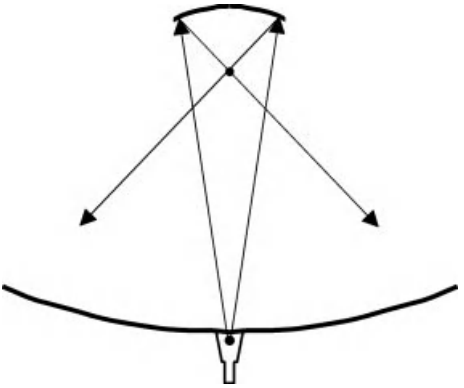


FIGURE 3.20 Center-fed Gregorian antenna with subreflector of same diameter as Cassegrain in Figure 3.19.

Every center-fed Cassegrain antenna is equivalent to a single-reflector paraboloidal antenna with the same feed but a greater *f*/*D* (Chang, 1989; Collin, 1985). The same is true of a center-fed Gregorian (Rusch et al., 1990). The properties of the equivalent paraboloid apply to the dual-reflector antenna (Chang, 1989). A dual-reflector antenna is more compact than the equivalent paraboloid, which is sometimes the main advantage.

Another one of the main advantages of the dual-reflector antenna over the single-reflector is that the former allows more control over the antenna's aperture field. The process for mutually shaping the main reflector's and subreflector's surfaces is, roughly speaking, to first shape the subreflector to control the aperture's power distribution and then to shape the main reflector to correct the aperture field's phase distribution (Collin, 1985).

An example of one main reflector that serves as both a single- and a dual-reflector, center-fed antenna is on NASA's Tracking and Data Relay Satellites (TDRSs). The "dichroic" subreflector of the Cassegrain reflects S-band radiation but allows K-band to pass. A proof-of-concept antenna built for the TDRS system (TDRSS) in 1977 is shown in Figure 3.21a, where the subreflector with Ku-band feed attached and the S-band feed at the reflector can be seen. The subreflector surface was made of an array of resonant crossed dipoles. Near the resonant frequency the surface became reflective, while at the lower S-band frequency the dipoles were electrically so short as to be almost invisible (Agrawal and Imbriale, 1979). A drawing of TDRS-7 with two of this type of antenna is shown in Figure 3.21b. TDRS-7 was the last of the first-generation satellites, launched from 1983 to 1995. Third-generation TDRSs with basically the same antenna are still being built today for launch in 2012 and 2013 (Boeing, 2008). These antennas are reconfigurable in that they are steerable.

Another example of a center-fed Cassegrain antenna is the V-band crosslink antenna on the US military's Milstar and Advanced Extremely High Frequency (AEHF) satellites. The antenna is reconfigurable in that it autotracks a signal

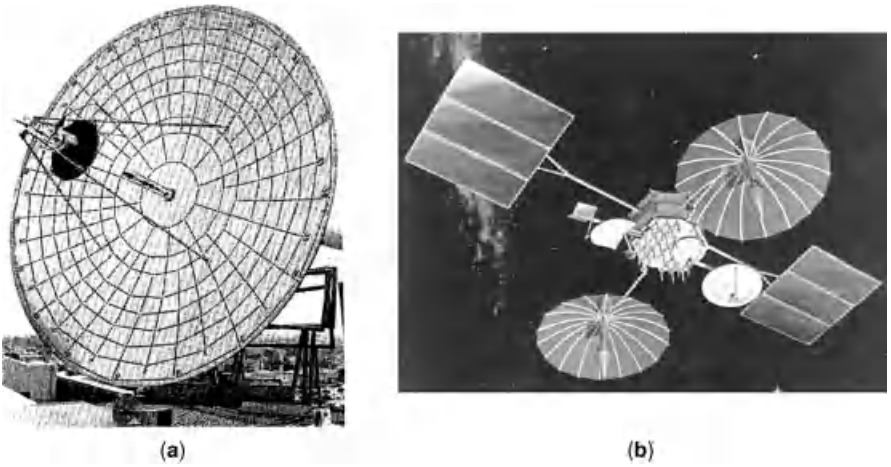


FIGURE 3.21 TDRS dual-frequency Cassegrain: (a) proof-of-concept antenna (©1979 IEEE. Reprinted, with permission, from Agrawal and Imbriale (1979)) and (b) TDRS-7 spacecraft with two such antennas (image courtesy of NASA).

(Choung, 2005), so the beam moves. Milstars were launched from 1994 to 2003, and the first AEHF was launched in August 2010.

3.3.3.2 Offset-Fed Like the single-reflector antenna, the dual-reflector antenna can be made in an offset-fed configuration, as illustrated in Figure 3.22.

Just as for a single-reflector antenna, the LP cross-pol is worse for an offset dual-reflector antenna than for a center-fed one. A f/D on the order of 1 and a horn with at least 10 dB aperture taper (Section 3.2.3) has little cross-pol. There is no cross-pol when the feed's offset angle is zero (Chu and Turrin, 1973). For a Gregorian, the cross-pol can be essentially eliminated by turning the subreflector a little (Akagawa and DiFonzo, 1979).

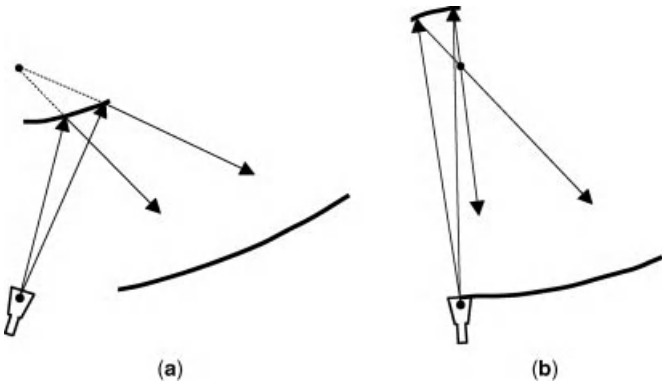


FIGURE 3.22 Offset-fed (a) Cassegrain and (b) Gregorian.

Every offset, classical Cassegrain or classical Gregorian antenna, even with a tilted subreflector, is equivalent to a single-reflector paraboloidal antenna with the same feed but a greater f/D. Assumptions are that the subreflector is at least 10–15 wavelengths in diameter and has an illumination edge taper of at least -10 dB (Rusch et al., 1990).

An example of an offset Gregorian Ku-band antenna is the steerable antenna on the Eutelsat W24 series of satellites, of which the first was launched in 1997 and the last a couple of years later. The antenna provides an elliptically shaped beam that can be fully reoriented by rotating the subreflector. Interestingly, the orientations of the dual LP stay fixed independent of beam orientation. Both reflectors are shaped. The subreflector is in the transition zone of the main reflector, that is, between its near and far fields. The feed is a corrugated horn (Pelaca, 1997).

Another example is the S-band antenna on the Sirius Radio satellites, the first three of which were launched in 2000 and the last remains on the ground as a spare. The antenna provides an elliptically shaped LHCP beam over CONUS.

3.4 HORN

3.4.1 Types of Horn

A horn can serve as an antenna in its own right when a relatively wide beam is desired or, much more commonly, as a feed for a reflector antenna or as a radiating element in an antenna array. A horn is a piece of waveguide which flares at the open end to increase the gain and improve the impedance match at the aperture. The horn can be rectangular or circular in cross-section. Important parameters of a horn antenna are its aperture efficiency and its polarization purity.

The **pyramidal horn** is a piece of rectangular waveguide which flares in both dimensions, usually by different factors. The flare angle must be rather small so that the phase across the aperture is almost constant. This horn can produce dual LP (Chang, 1989), but its cross-pol performance is not as good as that of the modified conical horns (see below).

A modified square horn with excellent cross-pol and high aperture efficiency has one or two steps in its flare (Section 3.5.4).

The **conical horn**, with circular cross-sections, is more common than the pyramidal horn. For such a horn, creating identical aperture distributions in the ***E***- and ***H***-planes can minimize the cross-pol (Bhattacharyya and Goyette, 2004). The simplest conical horn is smooth-walled with a linear taper, but its cross-pol is not good.

Various modifications to the simplest conical horn have been developed to improve the cross-pol. Perhaps the first development was the **Potter horn**, either smooth-walled or corrugated inside. It has a step between the waveguide and the flare, as shown in Figure 3.23, to excite the TM_{11} mode from the TE_{11} mode (Collin, 1985). **TM** is **transverse magnetic** and means that the magnetic field is perpendicular to the direction of propagation and the electric field is not (see

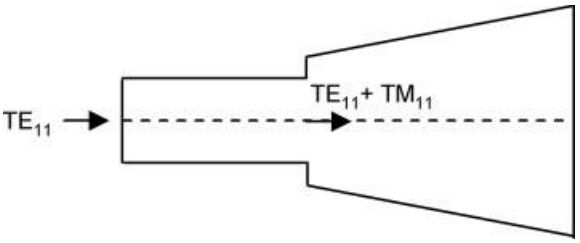


FIGURE 3.23 Potter horn (after Collin (1985)).

Section 4.3.2 for a discussion on waveguide propagation modes). The presence of the TM_{11} improves the ***E***-plane radiation characteristics (Evans, 1999). It is **dual-mode** because it has the two modes TE_{11} and TM_{11} at the aperture. The Potter horn has typically 70% aperture efficiency and not more since the aperture illumination is tapered to cause the magnetic field to go to zero on the horn wall, as the electric field does naturally (Bhattacharyya and Goyette, 2004).

Probably the most widely used conical horn is the **corrugated horn**, which has circular corrugations on the inner wall (Collin, 1985). It supports **hybrid modes**, which are combinations of TE and TM modes. The **dominant** hybrid mode HE_{11} is the combination of TE_{11} and TM_{11}. (The dominant mode is the one with the lowest operating frequency.) The horn can be designed to have more modes than just the dominant one in order to have a more uniform aperture distribution and higher beam efficiency, but it will then also have a narrower bandwidth (Chang, 1989). The horn can provide CP or LP, single or dual. It has a circularly symmetric pattern, very low cross-polarization, high beam efficiency, and low sidelobes (Chang, 1989).

A newer conical horn with very good cross-pol is smooth-walled but has a spline-curve tapering, which generates the necessary multimodes. This horn is lighter than the corrugated horn (reported in Schennum (2010a)).

3.4.2 Horn as Antenna

A horn antenna on a GEO satellite can provide a global beam, subtending at least $\pm 8.7^\circ$. An example of corrugated-horn antennas currently on orbit are the C-band global horns on the Intelsat-IX series of satellites launched from 2001 to 2003. The antennas are **dual-mode** (i.e., dual-hybrid-mode, different from the dual mode of the Potter horn) corrugated horns, with the second hybrid mode HE_{12} being excited at the second step, as sketched in Figure 3.24. The presence of the second mode flattens the aperture phase distribution, leading to an EOC gain increase of 0.4 dB and steeper gain roll-off beyond. One antenna receives both circular polarizations at once and the other transmits both at once. The port-to-port isolation of the antennas, inclusive of OMT and polarizer, is 37 dB and their cross-polarization isolation is 41 dB (Schennum et al., 1999).

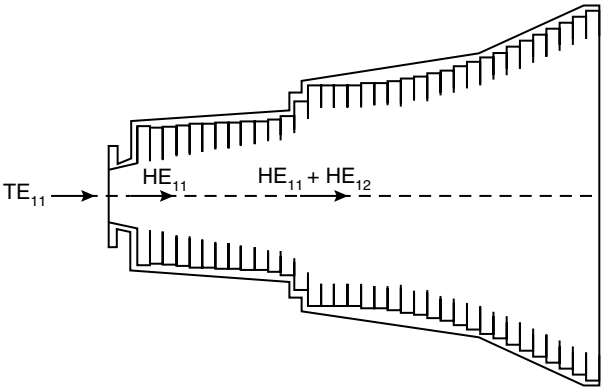


FIGURE 3.24 Dual-mode corrugated horn antenna on Intelsat-IX spacecraft. (©1999 IEEE. Reprinted, with permission, from Schennum et al. (1999)).

3.4.3 Horn as Feed for Single-Beam Reflector Antenna

For a reflector antenna creating one beam, these days the feed seems always to be a horn. The feed is a critical part of the antenna. The cross-pol of any reflector antenna can be no better than that of the feed (Collin, 1985). The feed can compensate some problems inherent in the reflector(s); for example, a corrugated-horn feed when excited with multiple modes can compensate the LP depolarization inherent in offset antennas, and provide dual LP (Adatia et al., 1981).

A corrugated-horn feed can have such low cross-pol levels as to allow communications on both polarizations. Some examples are as follows. An engineering-model horn, developed under contract from a large satellite operator, was able to receive at 6 GHz and to transmit at 4 GHz in dual CP (Tao et al., 1996). A flight horn that allows for a reconfigurable antenna is a Ku-band spot-beam feed. It can both receive and transmit on dual LP, whose orientation is partially reorientable, with cross-pol of −41 dB (SGM, 2000). Another such horn for a reconfigurable antenna with shaped reflector is a flight C-band feed. It can both receive and transmit on dual polarizations and can electromechanically switch between LP and CP. Its cross-pol is −37 dB or better for dual CP and −55 dB or better for dual LP (SGM, 2000).

A variation on the corrugated horn for dual CP was developed in the late 1990s by a satellite manufacturer. It has only one corrugation but a choke at the feed aperture, which together generate higher-order hybrid modes to make the $\boldsymbol{E}$- and $\boldsymbol{H}$-plane radiation patterns the same (Schennum and Skiver, 1997).

The phase center of the feed must be known so that it can be placed at the focus of the main reflector. A horn feed can be designed to have its phase center anywhere between the unflared waveguide and the aperture. However, the phase center shifts with frequency (Chang, 1989).

Horns used in an antenna array or a feed cluster for a reflector-based multibeam antenna are chosen on somewhat different criteria, as discussed in Sections 3.5.4 and 3.6.2, respectively.

3.5 ANTENNA ARRAY

An antenna array is an array of radiating elements of the same kind, laid out in a regular pattern. The outline of the array can be a circle, hexagon, or something less symmetric. The author could find no example of a satellite array being anything other than flat. The array forms beam(s) by combining radiating elements in either particular phase or particular delay relationships. One that uses phases is a phased array. One that uses delays is also sometimes said in the literature to be a phased array.

3.5.1 Array Principle

The basic principle of an array antenna is illustrated in Figure 3.25a, which shows a wavefront arriving in the direction of the arrow at the last two elements in a linear array (the linear array being used here only for explanatory purposes). The direction represents an angle θ off boresight. The dashed line represents the wavefront hitting the middle of the second-leftmost array element a short time ago and the solid line represents the wavefront hitting the middle of the leftmost element now. From θ and the distance between array elements, the distance d that the wavefront traveled between the two times can be computed. The delay t can be computed from d and the speed of light. Note that this delay is independent of frequency so applies to all frequencies. Delaying the earlier signal by t lines it up in time with the signal being received now, so the two can be added. The signals being received at the other elements can be correspondingly delayed and added in. Now, by the antenna reciprocity property, a signal could also be transmitted in the direction θ off boresight by delaying the individual signal copies going to the various array elements in a similar way.

A similar effect can be achieved by the use of phasing instead of delays, as illustrated in Figure 3.25b. Phasing is used in phased arrays. We have to think about things a little differently from in the basic delay principle. We think about the two different wavefronts that are hitting the middle of the same two array elements, now. The one hitting the second-leftmost element is phase α ahead of the one hitting the leftmost element. Phase difference α can be computed from d, the speed of light, and the center frequency f_0 of the signal being received. Note that α is not

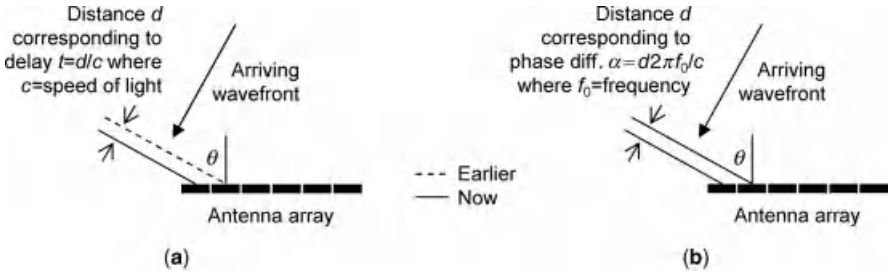


FIGURE 3.25 Principle of a linear array in terms of (a) time delay and (b) phase difference.

quite correct for any other frequency than f_0. Shifting the signal from the second-leftmost element by $-\alpha$ puts it in phase with the signal hitting the leftmost element, so the two can be added in-phase. The signals being received at the other elements can be correspondingly phase-shifted and added in. An antenna array that uses phases instead of delays is inherently bandlimited. However, most antenna arrays do use phases.

3.5.2 Antenna Array Characteristics

As discussed in Appendix 3.A.3, the antenna pattern of the array antenna is approximately given by the product of an individual radiating element's antenna pattern and the so-called **array factor**. The array factor is related to the spacing of the elements, among other things. When the spacing is the wavelength or less, the array's antenna pattern will typically have **grating lobes** (Kraus and Marhefka, 2003), also defined in the appendix, which are lower-level duplicates of the main pattern that have a regular repetition period in off-boresight angle. The closer together the elements are in the array, the larger the repetition period is, so the grating lobes are less of a problem. With large element spacing in the array, the grating lobes will send significant power in unwanted directions. Grating lobes arise because the array antenna's aperture as a whole consists of spatially discrete subapertures (from the radiating elements).

What makes the pattern differ from the product of the element pattern and the array factor is that the array elements show **mutual coupling**. The antenna pattern of the edge elements of the array is different from that of the interior elements. The result is that the array's sidelobes are not as low as they would be without mutual coupling (Chang, 1989).

A main beam that is off boresight is said to be **scanned**. Scanned beams have different properties from a boresight beam. For one thing, the antenna gain is less since gain is proportional to the apparent area of the array in the direction. For another, CP is no longer circular but elliptic—just imagine a circle being viewed off axis. Linear polarizations are no longer orthogonal, in general; the exception is if the scan is around an axis parallel to one of the linear polarizations. A third effect of scanning, for phased arrays, is that for all the signal frequencies except the nominal frequency, the beam has squint, that is, points not quite in the intended direction, as illustrated in Figure 3.26.

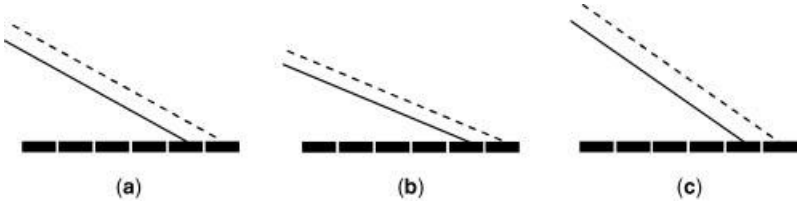


FIGURE 3.26 Wavefronts added in phase, for various frequencies, in a scanned beam (a) at nominal frequency, (b) at higher frequency, and (c) at lower frequency.

A design concern for satellites that create beams with overlapping coverages is the potential interference into any beam from the satellite's other beams. This concern applies to array antennas as well as to multibeam reflector antennas. Chapter 12 is devoted to the computation of the ratio of one beam's power to the interfering beams' power (*C/I*) and its variation on orbit.

3.5.3 Array Antenna

Array antennas are more often used as feeds for a reflector than as stand-alone antennas, but there are two current applications for the **direct-radiating** array antenna.

One application is on US government satellites. One example is on the TDRSs, their multiple-access receive and transmit phased array at S-band. The latest satellites launched, TDRS-H, I, and J, form five receive beams from 32 elements and one transmit beam from 15 elements (NASA, 2009). Another example is on the AEHF satellites. Each satellite has a 44 GHz receive phased array and two 20 GHz transmit-phased arrays (Northrop, 2007). A third US government satellite series with phased arrays is the Wideband Global SatCom satellites (WGS), sometimes called the "Wideband Gapfiller Satellites." The first block of the WGS satellites was launched from 2007 to 2009. A WGS satellite has a 188-element receive-phased array and a 312-element transmit-phased array, both creating up to eight beams at X-band (Nguyen et al., 2003). Receive and transmit are on opposite senses of CP. Each beam can have arbitrary shape and arbitrary size from 2.2° up to earth coverage (Roper et al., 2003). For all the satellites the beams are reconfigurable on orbit.

One reason for having direct-radiating phased arrays on military satellites such as Milstar and AEHF is that a larger number of radiating elements is possible than if the phased array served as a feed for a reflector MBA. Having a large number of radiating elements allows not only precise shaping of the coverage area but also good nulling, that is, very low gain in one or more directions, for example, toward a jammer. The weight and cost of a large phased array would be of secondary importance to the military.

The other application for the array antenna is on LEO satellites, such as the Iridium and Globalstar-2 spacecraft. The Iridium satellites were launched from 1991 to 2002. The first 18 Globalstar-2 spacecraft were launched in 2010 and 2011 and the remaining six are planned for 2012 (Globalstar, 2011). These satellites must create multiple beams to cover the earth below, which subtends ±55°. Figure 3.27 shows the two spacecraft. A guess of why phased arrays are feasible as antennas on LEO spacecraft but not on commercial GEO spacecraft is that the beams are bigger on the LEOs so the arrays are feasibly small. The beam angles are so far apart that the arrays would have to be replaced by many reflectors, which would be a mechanical problem.

The left-hand drawing shows an Iridium satellite, with its triangular earth deck having four gateway antennas on it and with three phased-array panels flaring out from the earth deck. Not readily visible in the drawing are the microwave crosslink antennas, which are in a ring between the earth deck and the phased-array panels.

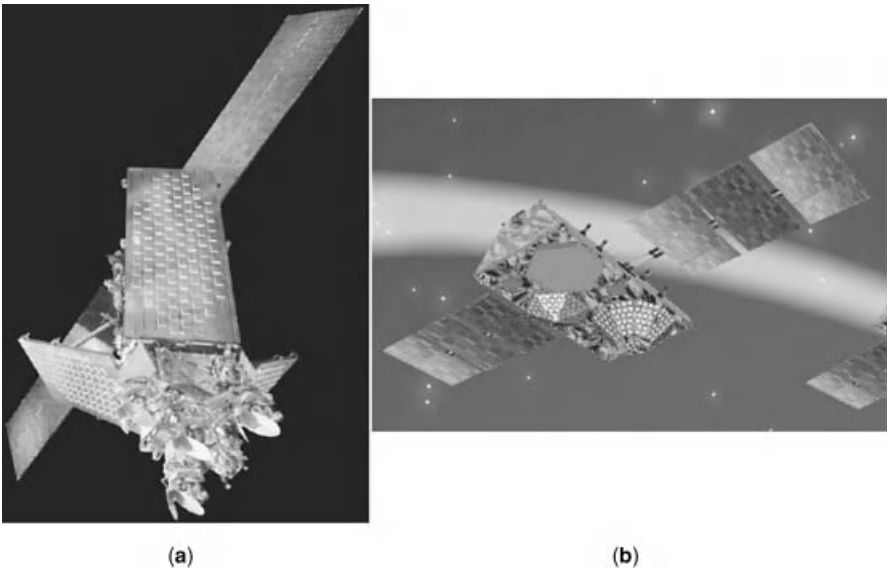


FIGURE 3.27 LEO satellites with phased-array antennas (a) Iridium (photo by Eric Long, National Air and Space Museum, Smithsonian Institution, SI 2006-935) and (b) Globalstar-2 (E. Briot © Thales Alenia Space).

The phased arrays transmit and receive at L-band. Each array panel has more than 100 elements and creates 16 fixed-size beams, which cover a one-third wedge of the earth below.

The right-hand drawing shows a Globalstar-2 satellite, with its two collections of transmit S-band phased arrays on the earth deck having shiny radiating elements, both collections having a flattened cone shape. The flat hexagonal shape on the earth deck is the receive phased array, at L-band. For both receive and transmit, the beams are the same 16: one in the middle about nadir, a ring around that beam of 6 beams, and an outer ring around it of 9 beams. The two transmit array collections together provide the 16 required beams. The larger array collection actually has a horn in the middle to create the nadir beam and 9 arrays around it that each create one of the beams in the outer ring. Each of these 9 is a phased array of 16 patches, tilted so its beam is on boresight. The smaller array collection has 6 arrays that each create one of the beams in the inner ring. Each of these 6 is a phased array of eight patches, which are also tilted (Lepeltier et al., 2007).

3.5.4 Array Radiating Elements

There are a few types of radiating elements that are used in antenna arrays.

One type of element is the helix in a cup, used on the first generation of TDRS for the S-band multiple-access antenna (NASA, 2009). A helix produces a beam of one sense of CP.

Another type of element is the microstrip patch, which consists of conducting material on a dielectric substrate above a ground plane. It is a printed circuit with dimensions on the order of half a wavelength. It can be made for frequencies between 1 and 10 GHz. Its bandwidth is small, usually only a few percent (Collin, 1985). It can be designed for LP or CP; it can be designed for two frequencies on the same LP sense or for dual LP (Chang, 1989). This type of array element is used to produce one sense of CP on the second generation of TDRS (NASA, 2009) and on the Iridium satellites (Schuss et al., 1999).

Another type is the annular slot made of strip transmission line. It is used to produce one sense of CP on the Globalstar satellites (Metzen, 2000).

The horn is a frequent choice for array element when the array scanning angles are low (Bhattacharyya and Goyette, 2004). For a given array size, the higher the aperture efficiency of each horn, the smaller the number of elements required, if the grating lobes do not become a problem (Section 3.5.2).

In the early 2000s, a satellite manufacturer developed high-efficiency horn elements with both circular and square apertures, suitable for use in phased arrays. The circular-aperture horn has LP cross-pol as low as that of a Potter horn but higher efficiency, 90% over 15% bandwidth. Its $\boldsymbol{E}$- and $\boldsymbol{H}$-plane aperture distributions are identical and almost uniform. The horn has smooth walls and two steps, one of which may be inward. The square-aperture horn has efficiency of 90% over 10% bandwidth. This horn has steps on the $\boldsymbol{H}$-plane for a single LP sense and steps on both planes for dual LP. For both horns, mutual coupling effects in an array can be neglected (Bhattacharyya and Goyette, 2004).

3.5.5 Passive and Active Arrays

Antenna arrays may be passive, active, or something in between. A passive antenna array has no amplifiers at the array elements; the amplifiers are in the repeater. In contrast, an active, transmit antenna array has a preamplifier and solid-state power amplifier (SSPA) at each array element. It may have the upconverters after the **beam-forming network** (**BFN**) (Section 3.5.6) and before the SSPAs, to obtain a decrease in losses between the BFN and the upconverters. Similarly, an active, receive antenna array has a low-noise amplifier (LNA) at each array element. It may also have the downconverters after the LNA and before the BFN (Maral and Bousquet, 2002). The antenna array that is neither passive nor active but something in between, the semiactive array, has active elements not right at the radiating elements but in the beam-forming network (Perrott and Griffin, 1991). It is described in Section 3.5.7. The various transmit architectures are illustrated in Figure 3.28; the receive architectures would be similar but reversed.

An example of a passive array is that on the Eutelsat-W2A spacecraft, launched in 2009. The array receives and transmits on dual CP (Lepeltier et al., 2007).

One advantage of an active array for transmit is that the losses between repeater and antenna come before the power amplifiers so the power amplifiers do not have to make up for them; for receive the corresponding advantage is that the losses between antenna and repeater come after the LNAs so they scarcely degrade the

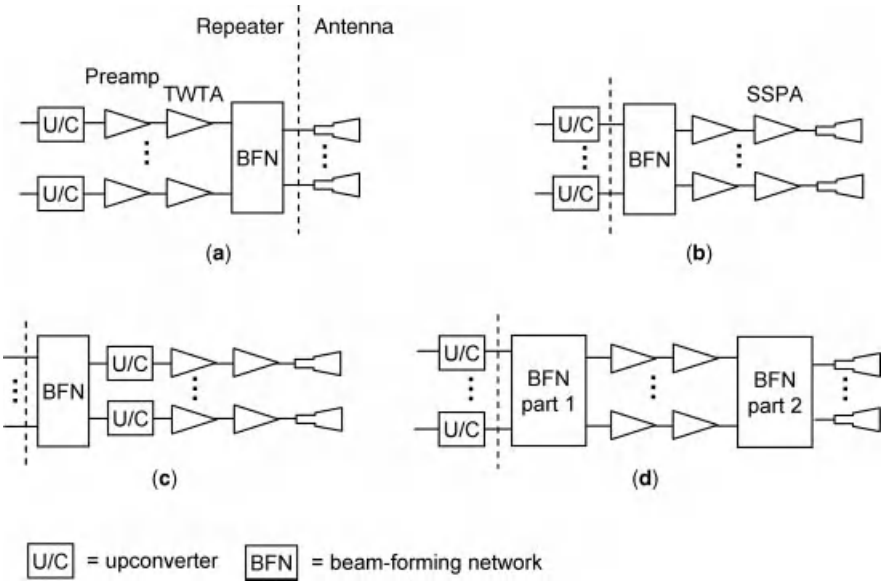


FIGURE 3.28 Transmit phased array architectures, also showing part of repeater: (a) passive, (b) active, (c) alternative active, and (d) semiactive.

payload noise figure. Another advantage of an active array is that redundant amplifiers are not needed because the array itself has redundancy. With some high probability, no more than some small portion of amplifiers will fail over the spacecraft lifetime. The array can be designed so that there are enough elements to begin with that no matter what array elements are lost the resultant amplification as well as the antenna pattern will be satisfactory, with high probability. Other advantages on transmit are that amplification by SSPA can be more linear than makes sense with TWTAs and that higher power per beam can be obtained than with TWTAs (Maral and Bousquet, 2002). A disadvantage on transmit is that SSPAs are less DC-power efficient than TWTAs. A big disadvantage of an active array is the greater number of units and the accompanying mass and mechanical issues.

Three examples of active arrays are the following:

- The receive array on the AEHF spacecraft. It is at 44 GHz and uses indium-phosphide semiconductors for the LNAs (Northrop, 2004).
- The receive and transmit arrays on the WGS spacecraft. The transmit array has amplitude taper, achieved by means of SSPAs at three power levels (Roper et al., 2003).
- The arrays on the Iridium spacecraft. Each of the more than 100 array elements has a transmit/receive module on its back side. The element is switchable between receive and transmit. The bias on the element's SSPA can be changed in order to reduce SSPA power draw during low-traffic times (Schuss et al., 1999).

3.5.6 Beam-Forming

The BFN is the part of an array antenna that creates the beams. There can be more or fewer beams than array elements or the same number. On receive, a BFN collects the signals from the various array elements, phases or delays them correctly, and adds them. On transmit, it power-divides the signal, phases or delays the resultant signals appropriately, and feeds them to the array elements. The beam can be shaped by applying different phases or delays, other than those required to point, and possibly different amplifications to different array elements. The BFN coefficients may have fixed values, for unchanging beams, or be resettable on orbit, for reconfigurable beams.

The BFN can be either on the spacecraft or on the ground. The advantage of beam-forming on the ground is that less onboard computing power is required, while a disadvantage can be the extra delay, more about which we will discuss below.

An example of on-the-ground beam-forming is the TDRSS. Since the beginning of the program the BFN has been on the ground. The satellite puts the signals from every one of the phased array's receive elements onto separate carriers and transmits them to the ground. The BFN there combines all the signals with proper phases and amplitudes to create each beam, using a different set of amplitudes and phases for each beam. Similarly, on the ground the element phases for the transmit beam are computed and all the element signals phase-shifted. The signals are all sent up to the satellite on separate frequencies, where they drive all the elements. The beams are scheduled (Zillig et al., 1998).

The only other example of beam-forming on the ground known to the author is the TerreStar-1 satellite, launched in 2009, which can form more than 500 beams (Epstein, 2009). Onboard beam-forming would be prohibitive.

Onboard beam-forming has several varieties. Fixed beams come from fixed phase shifters or fixed delay elements in the BFN. Reconfigurable beams are created by either **electronic beam-steering** or **digital beam-forming**. In the former, typically there are phase shifters and sometimes amplitude controls, and these are electronically commandable to either a fixed, limited set of values or to a continuous set of values within the available resolution. In the latter, all computations are performed digitally and delays instead of phases are used; the payload is considered to be a processing payload (Section 9.4.4).

Examples of satellites with onboard beam-forming are the following:

- *WGS Satellites:* Both the receive and transmit phased arrays have both phase and amplitude control (Nguyen et al., 2003).
- *Iridium Satellites:* The phased arrays, each used for both receive and transmit, have both phase and amplitude control (Schuss et al., 1999).
- *Spaceway-3:* The downlink phased array of 1500 elements creates 24 simultaneous downlink **hopping spot beams**, each of which hops among 32 beams (Hughes, 2010). The hopping rate is unknown to the author, but it is conceivably comparable to the 40,000 hops per second demonstrated on NASA's

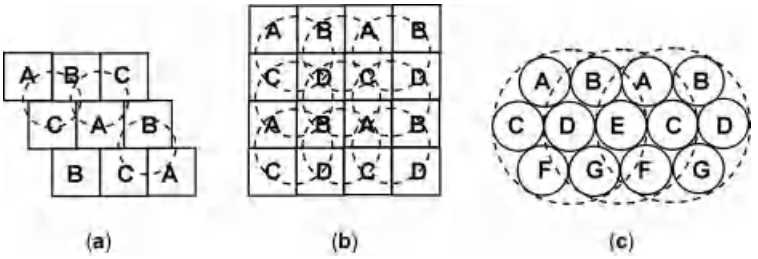


FIGURE 3.29 Phased-array element clusters forming overlapping feeds: (a) three elements per feed, (b) four elements per feed, and (c) seven elements per feed. (©1990 IEEE. Reprinted, with permission, from Roederer and Sabbadini (1990)).

ACTS, launched in 1993 (NASA, 2000), or to the 20,000 hops per second of a phased-array antenna reported in 2000 to be under development by a space manufacturer for government applications (Warshowsky et al., 2000).

3.5.7 Semiactive Array with Multimatrix Amplifier

The author found only one case of a phased array which is **semiactive**, meaning that there are amplifiers in the near vicinity of the radiating elements but they are not one-to-one with the elements. The architecture for this is provided by "multimatrix amplifiers" (Section 6.5.1.1) being incorporated into the BFN, as shown at a high level in Figure 3.28d. This case is also the only one found of an array antenna which does not use all radiating elements for every beam. These two things go hand in hand. Because the BFN does not need to connect all elements for every beam, the BFN is smaller than usual but conceptually more complicated (Perrott and Griffin, 1991). ESA patented this approach (Nubla et al., 1998). All amplifiers are driven at the same level, for efficiency, but the beams have different powers. The architecture can be used for receive or transmit (Perrott and Griffin, 1991).

The phased array forms adjacent beams from small clusters of overlapping elements. The concept is illustrated in Figure 3.29, which shows examples of beams formed from three, four, and seven elements (Roederer and Sabbadini, 1990).

3.6 REFLECTOR-BASED MULTIBEAM ANTENNA

3.6.1 Reflector MBA Concepts

This section deals with the reflector-based **multibeam antenna** (**MBA**). Such an antenna has a single or dual reflector, and its feeds simultaneously create multiple beams. There are three kinds of feeds that a MBA can have, each of which will be discussed in a separate subsection below:

- Feed cluster, that is, cluster of horns, each horn a feed for one beam
- Feeds formed by an antenna array
- Overlapping feeds formed from phased, overlapping clusters of radiating elements.

Offset feeding seems to be the rule for reflector MBAs, as it has several big advantages. For one, the multiple feeds will not block the aperture (Rudge et al., 1982). Having a larger *f*/*D* has at least two good effects: the feeds will be larger and so will have reduced mutual coupling (Rudge et al., 1982), and antenna performance is less sensitive to the feeds not all being exactly at the focal point (Howard, 1983).

No matter what kind of feeds the MBA has, the electrical fields of the radiating elements interact with each other, causing mutual coupling. For example, both the co-pol and cross-pol characteristics of a conical horn change when it is in a cluster, even when the horn in isolation has excellent cross-pol (Schennum and Skiver, 1997). The result of mutual coupling is that the antenna pattern of elements in the middle is in general different from that of the elements on the edge of the group (Collin, 1985).

A MBA beam is scanned if its direction is off boresight. All feeds besides the one exactly at the focal point create scanned beams. There has been considerable study over the years on the properties of scanned beams from offset reflector antennas. The offset axis can be defined as the line originating at the ideal feed location and pointing to the middle of the first reflector that it sees. The offset angle is the angle between negative boresight and the offset axis. The offset **focal plane** is the plane perpendicular to the offset axis and containing the ideal feed location (Ingerson and Wong, 1974). For a Gregorian MBA with subreflector turned a little as discussed in Section 3.3.3.2 and feed displaced a little in the offset focal plane, the gain of the Gregorian is not well predicted by the gain of the equivalent single-reflector antenna, although the scanning direction is (Akagawa and DiFonzo, 1979). A similar statement holds for a Cassegrain with large scans, but with small scans the model predicts well (Rahmat-Samii and Galindo-Israel, 1981). The optimum surface to locate displaced feeds on is not the offset focal plane but a more complicated surface (Krichevsky and DiFonzo, 1982; Sletten and Shore, 1982). However, all MBA feeds that the author has seen lie in a plane.

3.6.2 Reflector MBA with Feed Cluster

A **feed cluster** is a collection of separate feeds, each creating a different beam. The feeds in a cluster seem always to be horns. Examples of feed horns for a MBA are the following:

- A Ka-band flight feed that can both receive and transmit on dual CP with cross-pol of −34 dB (SGM, 2000).
- High-efficiency feed horns developed by a satellite manufacturer in the early 2000s, for single or dual LP. The circular-aperture horn has cross-pol as low as that of a Potter horn but a higher efficiency. The square-aperture horn has the

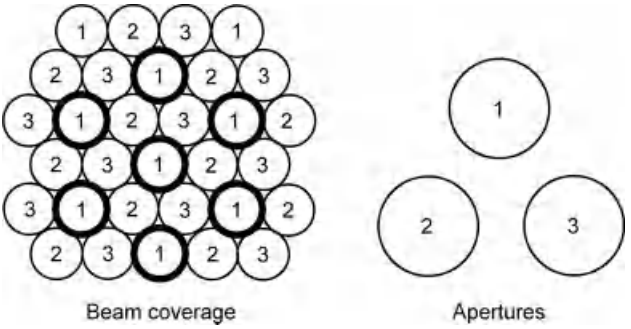


FIGURE 3.30 Three-aperture MBA design, where the numbers represent frequency cells (after Rao (1999)).

same efficiency but over a narrower bandwidth. (See Section 3.5.4 for more information on these horns.)

- High-efficiency feed horns flown on Anik F2 for dual CP and simultaneously 19 and 29 GHz. Anik F2 was launched in 2004. These smooth-walled conical horns have very good cross-pol (Amyotte et al., 2006).

High-efficiency feeds are not always the best choice—depending on the antenna geometry, the beam layout, and the frequency reuse scheme, sometimes lower-efficiency feeds are better for the overall MBA performance (Bhattacharyya and Goyette, 2004).

If the coverage areas of the beams are to be in a hexagonal pattern that completely covers a large area, the feeds would have to be unacceptably limited in size if they are all in the same feed cluster. Instead, having three or four reflectors on the spacecraft allows beams with overlapping, adjacent coverages to be in different feed clusters and therefore to have larger apertures (Rao, 1999). This is illustrated in Figures 3.30 and 3.31, respectively. A larger aperture means higher beam gain, sharper gain roll-off at beam edge, and lower sidelobes to interfere into other

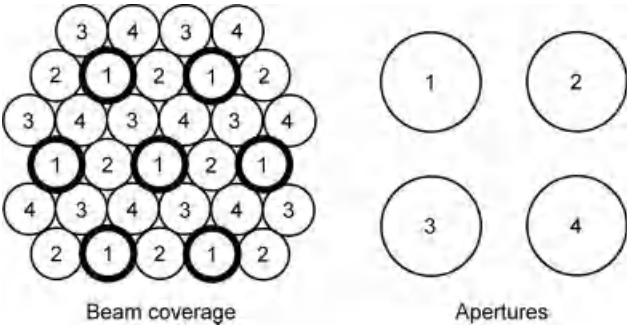


FIGURE 3.31 Four-aperture MBA design, where the numbers represent frequency cells (after Rao (1999)).

beams. The *f*/*D* of one of these multiple reflectors is typically between 1.0 and 2.0 (Rao, 2003).

A satellite that has three single-reflector MBAs each with a cluster of feed horns is the Ciel-2 satellite. The antennas operate in CP at Ku-band. The reflectors have shaped surfaces (Lepeltier et al., 2007).

3.6.3 Reflector MBA with Array Feeds

At GEO, the main advantage of using an antenna array with a reflector instead of allowing the array to directly radiate is that the array can be much smaller (Jamnejad, 1992).

One difficulty of using an antenna array for the feed, when it is desired to scan the beam, is that a 1° scan of the antenna corresponds to an $x°$ scan of the array, where x is the ratio of the reflector aperture diameter to the array diameter (Jamnejad, 1992). So the decreased gain and inferior CP of the scanned array come into play even at small antenna scan angles.

Some types of radiating elements are not suitable for array feeds at Ka-band because such an array has a very ill-defined phase center (Jamnejad, 1992).

At L-band, a phased array of 120 cup-helix radiators with a single reflector is used on the Inmarsat-4 series of satellites, launched from 2005 to 2008, and on the follow-on Alphasat-XL satellite to be launched in 2012 (Dallaire et al., 2009). Each cup-helix radiator is mounted on a large ground plane and has its own diplexer. They all provide the same sense of CP. The mutual coupling between elements is significant, even with the cups to isolate the elements. The MBA is reconfigurable in an unusual sense: the beam coverages are fixed, but since the satellites' inclination is allowed to slowly vary within $\pm 3°$ over the satellite's lifetime (Martin et al., 2007), the weights of the radiating elements are frequently (daily) modified (Guy et al., 2003).

An example at S-band is TerreStar-1, which has a phased array producing the feeds for its 18 m reflector, the largest commercial reflector flown by the time of its launch (Epstein, 2009). The satellite is depicted in Figure 3.32.

3.6.4 Reflector MBA with Overlapping Feeds

The author found only one satellite with phased arrays that form overlapping feeds for MBAs, the Inmarsat-3 series of satellites (Perrott and Griffin, 1991). These satellites were launched from 1996 to 1998. The MBAs each use a single reflector. Their semiactive BFNs and the concept of overlapping feeds were discussed in Section 3.5.7. The satellite has a receive MBA and a transmit MBA, the latter being a slightly scaled-up version of the former and both being at L-band. The satellite is reconfigurable because each has to be able to operate in any of five orbital slots once on orbit. This means it has to be able to create, at each orbital location, up to 7 out of a total of 22 possible beams. By means of the feed array being rotatable about boresight, the 22 are reduced to 7. Each antenna has 22 radiating elements in a

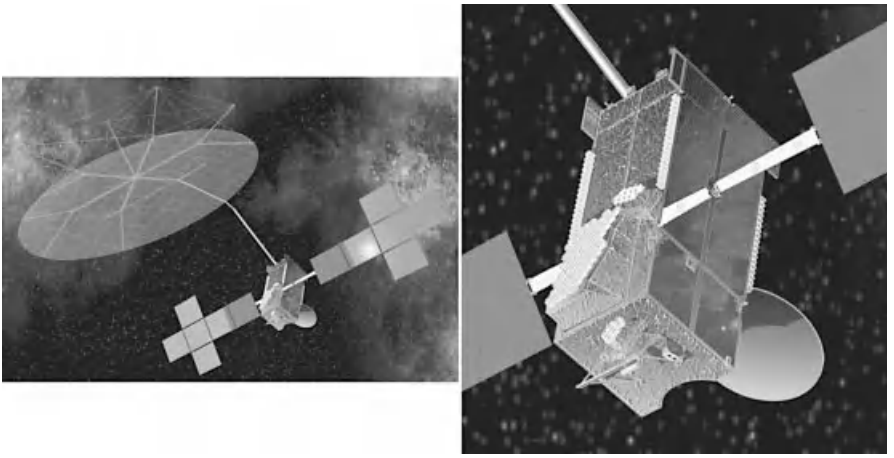


FIGURE 3.32 Reflector MBA: TerreStar-1 with phased-array feeds (image courtesy of Space Systems/Loral).

double ring with 1 in the center, and 5 elements are needed to form most beams (Perrott and Griffin, 1991).

The advantage of this type of MBA feed is that it takes only one reflector to create densely packed beams. In comparison, a feed cluster needs three or four (Section 3.6.2).

3.7 AUTOTRACK

Sometimes an antenna needs to point with more accuracy than the spacecraft attitude-control system alone provides. This is often the case with a multispot-beam antenna. Then the antenna will have an onboard **autotrack** system, which keeps the antenna's peak of beam pointed to a beacon or a communications signal (EMS, 2002). The system is almost always a closed-loop (= feedback) control system, which measures the pointing error and feeds back the negative of the error to the antenna steering mechanism as a control signal. The pointing error has two orthogonal components, azimuth and elevation.

More precisely, in the case of a reflector antenna, a tracking feed provides two "difference signals," one for azimuth error and one for elevation error. The difference signals are approximately proportional to the respective components of the pointing error, over the tracking range. The two angles are ideally orthogonal over the tracking range but in fact are never quite (see, e.g., Mahadevan et al., 2004). A tracking receiver takes these signals and creates the control signals, which are then input to the antenna steering mechanism.

Most often the autotrack system is of the monopulse variety, where the term "monopulse" comes from radar. A monopulse system has the following characteristics: (1) the difference signals can be continually measured; (2) the whole-signal

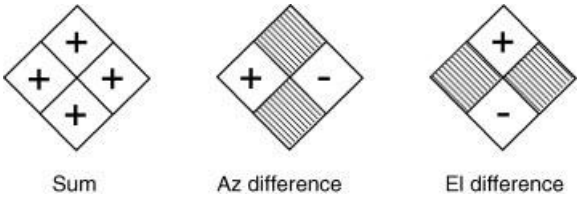


FIGURE 3.33 Four-horn diamond feed for tracking and communications: how sum and difference signals are formed from individual signals (after Skolnik (1970)).

(the "sum-signal") strength can also be continually measured, and it is used by the tracking receiver to scale the difference signals; and (3) the antenna steering system physically turns the antenna, either the reflector, subreflector, or the feed array (Howley et al., 2008). A good description of monopulse tracking is given in Skolnik (1970) and its later editions. A slight variation on the monopulse system is the pseudo-monopulse one, which electronically scans the beam instead of moving the antenna (Howley et al., 2008).

The tracking feed is located at the focal point of the reflector or subreflector. The feed senses any lateral displacement of the focal point from the middle of the feed. The feed provides measurement of the difference signals based on amplitude differences (Skolnik, 1970). There are two kinds of tracking feed.

In one kind, easier to understand, the feed consists of multiple horns packed together. Figure 3.33 gives an example with four horns, in which each pair of opposite horns measures the difference signal along the line that joins them and all the horns together receive the sum signal. When the antenna is perfectly pointed, in each pair of horns both horns measure the same amplitude, so their difference signal is zero. When the focal point is offset, in at least one pair of horns one of the horns will receive greater amplitude than the other will, so the difference signal will be nonzero.

The other kind of tracking feed is a single multimode horn. The aperture can be square (Skolnik, 1970) or round (Yodokawa and Hamada, 1981). When the focal point is in the middle of the feed, only the conventional waveguide mode of the feed is excited, by the sum signal. When the focal point is offset, higher-order waveguide modes are additionally excited and coupled off. Their magnitudes and signs give the orthogonal difference signals. The multimode horn is wider-band than the multihorn feed (Skolnik, 1970).

Both kinds of tracking feeds can perform dual LP or dual CP (Skolnik, 1970, for the four-horn diamond feed; Prata et al., 1985, for conical multimode) with the proper hardware behind them.

Typical total antenna pointing error is generally considered to be about 0.15° without autotrack and about 0.05° with autotrack (Goodzeit and Weigl, 2010). One antenna manufacturer's Ka-band autotrack system on a multispot-beam antenna contributed less than 0.01° to a total antenna pointing error of less than 0.04° (Amyotte et al., 2006). A patent recently awarded to a satellite manufacturer would cut the total pointing error to maybe half, by adding to the autotrack control loop a

feedforward component with a pointing prediction based on knowledge the system has gained on previous on-orbit days (Goodzeit and Weigl, 2010).

If the composite coverage is zonal or smaller, one beacon inside the coverage area may be enough. The closer the beacon is to the center of the coverage area, the better.

Inmarsat-4, though, has a larger, hemispherical coverage area. It uses two beacons, at the edge of the visible earth and about 90° apart. Four payload beams, ±1° from the beacon direction, are used for monopulse autotracking (Stirland and Brain, 2006).

APPENDIX 3.A

3.A.1 Decibel

Electrical engineers usually give ratios in **decibels (dB)**. This is a logarithmic scale. Communications-satellite power and noise levels have a wide range, often several orders of magnitude. Using dB flattens the range to something more humanly graspable. A ratio of powers P_1 and P_2 is given in dB as follows:

$$\text{Power ratio } \frac{P_1}{P_2} \text{ in dB} = 10 \log_{10}\left(\frac{P_1}{P_2}\right)$$

The factor of 10 in front of the logarithm causes the range of P_1/P_2 between 1 and 10 to have a dB range of 0–10, that is, almost the same, instead of the small range, without the factor of 10, of 0–1.

It is wise to memorize a few power ratios in dB, as shown in Table 3.1. Then one can calculate most other power ratios mentally by the various relationships given below:

$$10 \log_{10}(UV) = 10 \log_{10} U + 10 \log_{10} V \qquad 10 \log_{10} \frac{1}{U} = -10 \log_{10} U$$

$$10 \log_{10} \sqrt[n]{V} = \frac{1}{n} 10 \log_{10} V$$

TABLE 3.1 A Few Power Ratios in dB to Memorize

Numerical Power Ratio	dB Power Ratio
1	0
2	3.0
3	4.8
4	6.0
5	7.0
7	8.5
10	10
100	20
1000	30

For example, 30 in dB is $10\log_{10}(10 \times 3)$, namely about $10 + 4.8 = 14.8$ dB, and 15 in dB is about $14.8 - 3.0 = 11.8$ dB.

The term "numerical ratio" is frequently used to mean a ratio that is not in dB. In this book, the term more often used is "not in dB."

A ratio of amplitudes or voltages has a different definition of dB, and this can be confusing. The ratio of amplitudes A_1 and A_2 is given in dB as follows:

$$\text{Amplitude ratio } \frac{A_1}{A_2} \text{ in dB} = 20\log_{10}\left(\frac{A_1}{A_2}\right)$$

In fact, this is the same definition, since amplitude is proportional to the square root of power:

$$\text{Power ratio } \frac{A_1^2}{A_2^2} \text{ in dB} = 10\log_{10}\left(\frac{A_1^2}{A_2^2}\right)$$

What can also be confusing is that sometimes not just ratios but terms with units are given in a dB-type format, such as a frequency in dBHz. The question is, which definition of dB to use, the one with a factor of 10 on the log or the one with a factor of 20. The factor of 10 is used if, in a commonly used expression, the term either directly multiplies or divides a power. By "directly" we mean without taking its square or square root. An example is a bandwidth B, since it is in the common expression $\text{SNR} = P/(N_0B)$. Similarly, the factor of 10 is also used if, in a commonly used equation, the term is set equal to an expression in which a power or inverse power is directly present, as for the symbol energy E_s in the equation $E_s = P/R_s$. In fact, the definition with factor of 20 is not often used.

3.A.2 Antenna Pattern of General Aperture

The Fourier transform $G(f)$ of a time function $g(t)$ is defined as follows:

$$G(f) = \int_{-\infty}^{\infty} g(t)e^{-j2\pi ft}dt \quad \text{for all frequencies } f$$

There are some conditions that $g(t)$ must satisfy for it to have a Fourier transform (Cooper and McGillem, 1967), but for us they are benign. The **inverse Fourier transform** $g(t)$ of $G(f)$ is as follows:

$$g(t) = \int_{-\infty}^{\infty} G(f)e^{j2\pi ft}df \quad \text{for all times } t$$

Two qualitative properties of Fourier transform pairs are given in Figure 3.34, where "$\Leftrightarrow$" means "if and only if" and "$\Rightarrow$" means "implies."

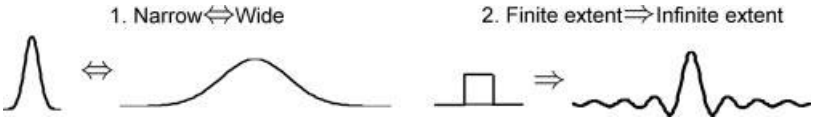


FIGURE 3.A.1 Properties of Fourier transform pairs, with examples.

Now, the far-field antenna pattern of a high-gain aperture antenna is very closely approximated near boresight by the two-dimensional Fourier transform of the tangential electric field in the aperture. (In any direction the far-field pattern is simply related to the two-dimensional transform, but away from boresight the relationship with the azimuth and elevation angles off boresight is not simple.) The two-dimensional Fourier transform $G(x',y')$ of a function $g(x,y)$ is given as follows:

$$G(x', y') = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) e^{-j2\pi(x'x+y'y)} dxdy$$

The preceding came from Collin (1985).

We can draw some lessons about antenna patterns. First, from Figure 3.A.1 we know that

1. In general the smaller the aperture is, the wider the pattern, for a fixed frequency.
2. The fact that an antenna has a finite extent means that its pattern will have not only a mainlobe but also sidelobes.

For a further application, suppose that the aperture is wider in one dimension than the other. Imagine a plane that cuts the aperture in the wide dimension and goes through the aperture center. Also imagine the plane for a cut in the short direction. The antenna pattern in the first plane is narrower than the antenna pattern in the second plane.

3.A.3 Antenna Pattern of Antenna Array

Suppose an antenna array's radiating elements all have the same tangential electric field in their apertures except for either different phases or different delays and possibly different powers. Assume also that the elements all have the same spatial orientation. (These both seem to always be the case.) A phased array is the spatial convolution of an individual aperture electric field with an array of ideal point sources located at the element locations and having different phases and powers. A time-delay array is similar except that the point sources have different delays instead of phases.

When the array is rectangular in shape and the elements could be placed on a rectangular grid, the array factor is the product of the two one-dimensional array factors (Collin, 1985).

Now, the two-dimensional Fourier transform of the convolution will be approximately equal to the array's antenna pattern, since the element is an aperture antenna

(Appendix 3.A.2). The equality is only approximate due to mutual coupling among the array elements. The closer together the array elements are and the smaller they are, the more mutual coupling there is. Now, the transform of the convolution is the product of the transforms, as is well known (Cooper and McGillem, 1967). The transform of the aperture field of one element is the antenna pattern of one element, and the transform of the point-source array is the array factor (Collin, 1985). Since the pattern of an individual element is in general very broad, the antenna pattern is mostly determined by the array factor (Collin, 1985).

Because the point-source array is a sampled function in space, its Fourier transform, that is, its antenna pattern, will repeat (Cooper and McGillem, 1967). However, it will not be exactly a repetition—the duplicates will have smaller gain than the main pattern because of the roll-off in the element's antenna pattern (Chang, 1989). These smaller duplicates are the grating lobes. The angular repetition period is larger the closer together the array elements are (Cooper and McGillem, 1967); it is desired that the grating lobes not be too close to the main pattern.

REFERENCES

Adatia N, Watson B, and Ghosh S (1981). Dual polarized elliptical beam antenna for satellite application. *IEEE Antennas and Propagation Society International Symposium;* 19; (June); 488–491.

Agrawal VD and Imbriale WA (1979). Design of a dichroic Cassegrain subreflector. *IEEE Transactions on Antennas and Propagation;* 27; 466–473.

Akagawa M and DiFonzo DF (1979). Beam scanning characteristics of offset Gregorian antennas. *IEEE Antennas and Propagation Society International Symposium;* 17; (June); 262–265.

Albertsen NC and Pontoppidan K (1984). Analysis of subreflectors for dual reflector antennas. *IEE Proceedings, Part H: Microwaves, Optics and Antennas;* 131; (June); 205–213.

Amyotte E, Demers Y, Martins-Camelo L, Brand Y, Liang A, Uher J, Carrier G, and Langevin J-P (2006). High performance communications and tracking multi-beam antennas. *Proceedings of European Conference on Antennas and Propagation;* Nov. 6–10; 1–8.

Bhattacharyya AK and Goyette G (2004). A novel horn radiator with high aperture efficiency and low cross-polarization and applications in arrays and multi-beam reflector antennas. *IEEE Transactions on Antennas and Propagation;* 52; 2850–2859.

Boeing Defense, Space & Security (2001). DIRECTV uses spot beam technology on DIRECTV 4S. Product sheet. Nov. On www.boeing.com/defense-space/space/bss/factsheets/601/directv_4s/directv_4s.html. Accessed Oct. 2011.

Boeing Defense, Space & Security (2008). TDRS K, L. Program fact sheet on www.boeing.com/defense-space/space/bss/programs.html. Accessed Jan. 12, 2010.

Chang K, editor (1989). *Handbook of Microwave and Optical Components, Vol. 1, Microwave Passive and Antenna Components*, New York: John Wiley & Sons, Inc.

Cho SH, Parekh SV, and Profera CE (1985). Multi-beam offset Cassegrain antenna for satellite communication applications. *IEEE Antennas and Propagation Society International Symposium;* 23; (June); 605–608.

Choung Y (1996). Dual-band offset gimballed reflector antenna. *Digest, IEEE Antennas and Propagation Society International Symposium;* 1 (Jul. 21–26); 214–217.

Choung YH (2005). V-band cross-link antenna. *IEEE Antennas and Propagation Society International Symposium;* 3A; (Jul. 3–8); 387–390.

Chu T-S and Turrin RH (1973). Depolarization properties of offset reflector antennas. *IEEE Transactions on Antennas and Propagation;* 21; 339–345.

Collin RE (1985). *Antennas and Radiowave Propagation*, New York: McGraw-Hill, Inc.

Cooper GR and McGillem CD (1967). *Methods of Signal and System Analysis*, New York: Holt, Rinehart and Winston, Inc.

Dallaire J, Senechal G, and Richard S (2009). The Alphasat-XL antenna feed array. *Proceedings of European Conference on Antennas and Propagation;* Mar. 23–27; 585–588.

Duret G, Guillemin T, and Carriere R (1989). The EUTELSAT II reconfigurable multi-beam antenna subsystem. *Digest, IEEE Antennas and Propagation Society International Symposium;* 1; (June 26–30); 476–479.

EMS Technologies Inc (2002). Autotrack combiners. Application note 44X-1. June. On www.emsdss.com/uploadedFiles/pdf/ATM.pdf. Accessed Feb. 28, 2011.

Epstein J, TerreStar Corp (2009). TerreStar—Year in review. SatMagazine; Dec.; p. 156.

Evans BG, editor (1999). *Satellite Communication Systems*, 3rd ed., London: The Institution of Electrical Engineers.

Gagliardi R (1978). *Introduction to Communications Engineering*, New York: John Wiley & Sons., Inc.

Globalstar, Inc (2011). Globalstar announces December 25 date for third launch of six new second-generation satellites. Press release. Nov. 29. On www.globalstar.com. Accessed Dec. 1, 2011.

Goodzeit NE and Weigl HJ, inventors; Lockheed Martin Corp, assignee (2010). Antenna autotrack control system for precision spot beam pointing control. U.S. patent 7,663,542 B1. Feb. 16.

Guy RF, Wyllie CB, and Brain JR (2003). Synthesis of the Inmarsat 4 multibeam mobile antenna. *Conference Publication 491, International Conference on Antennas and Propagation;* 1 (Mar. 31–Apr. 3); 90–93.

Hoffmeister R, Space Systems/Loral (2010). Private communication, Mar. 5.

Howard HT, inventor; Chaparral Communications, Inc, assignee (1983). Feed horn for reflector antennae. U.S. patent 4,380,014. Apr. 12.

Howley RJ, Daffron WC, Hemlinger SJ, and Gianatasio AJ, inventors; Harris Corp, assignee (2008). Monopulse antenna tracking and direction finding of multiple sources. U.S. patent application publication 2008/0122683 A1. May 29.

Hughes Network Services, LLC (2010). Spaceway 3 technical specifications. On www.hughes.com. Accessed Mar. 2010.

Ingerson PG and Wong WC (1974). Focal region characteristics of offset fed reflectors. *IEEE Antennas and Propagation Society International Symposium;* 12; (June); 121–123.

Intelsat (2008). Intelsat satellite guide. On www.intelsat.com/network/satellite. Accessed Jan. 2010.

Jamnejad V (1992). Ka-band feed arrays for spacecraft reflector antennas with limited scan capability—an overview. *Digest, IEEE Aerospace Applications Conference;* Feb. 2–7; 57–66.

Kraus JD and Marhefka RJ (2003). *Antennas for All Applications*, 3rd ed., International ed. Singapore: McGraw-Hill Education (Asia).

Krichevsky V and DiFonzo DF (1982). Beam scanning in the offset Gregorian antenna. *Comsat Technical Review;* 12; (fall); 251–269.

Kwok RS and Fiedziuszko SJ (1996). Advanced filter technology in communications satellite systems. *Proceedings of International Conference on Circuits and System Sciences;* (June 20–25); 1–4.

Legay H, Croq F, and Rostan T (2000). Analysis, design and measurements on an active focal array fed reflector. *Proceedings of IEEE International Conference on Phased Array Systems and Technology;* May 21–25; 399–402.

Lepeltier P, Maurel J, Labourdette C, Croq F, Navarre G, and David JF (2007). Thales Alenia Space France antennas: recent achievements and future trends for telecommunications. *European Conference on Antennas and Propagation;* Nov. 11–16; 1–5.

Mahadevan K, Ghosh S, Nguyen B, and Schennum G (2004). TX, RX & $\Delta x/\Delta y$-autotrack CP feed for multiple beam offset reflector antennas. *AIAA International Communications Satellite Systems Conference;* May 9–12; 1–10.

Maral G and Bousquet M (2002). *Satellite Communications Systems*, 4th ed., Chichester, England: John Wiley & Sons Ltd.

Martin DH, Anderson PR, and Bartamian L (2007). *Communication Satellites*, 5th ed., El Segundo, CA; Reston (VA): The Aerospace Press; American Institute of Aeronautics and Astronautics, Inc.

Maufroid X, Coromina F, Folio B-M, Hughes R, Couchman A, Stirland S, and Joly F (2004). Next generation of transparent processors for broadband satellite access networks. *AIAA International Communications Satellite Systems Conference;* May 9–12; 1–13.

Metzen PL (2000). Globalstar satellite phased array antennas. *Proceedings of IEEE International Conference on Phased Array Systems and Technology;* May 21–25; 207–210.

NASA (2000). For ACTS experiments, all good things must come to an end. News release. May 30. On acts.grc.nasa.gov/news/acts-good-things-end.shtml. Accessed Mar. 10, 2010.

NASA (2009). TDRSS space segment. On www.spacecomm.nasa.gov/spacecomm/programs. Accessed Nov. 24, 2009.

Nguyen LN, Kinal GV, and Paul HI (2003). Performance of hybrid link power control using Wideband Gapfiller Satellite-based automatic level control (ALC) and ground-based link power control (LPC). *IEEE Military Communications Conference;* 1; (Oct. 13–16); 511–516.

Northrop Grumman Corp. (2004). Northrop Grumman continues advanced EHF payload successes by demonstrating uplink phased array antenna technology. News release. Sept. 28.

Northrop Grumman Corp (2007). Northrop Grumman integrates all phased array antennas on first Advanced EHF flight payload. News release. Jan. 23.

Nubla A, Montero J-M, and Bustamante L (1998). A high power hybrid matrix for mobile satellite communications. *Preparing for the Future*. ESA publication; Vol. 8; June

Pelaca E (1997). Ku-band elliptical steerable and rotatable beam antennas. *Digest, IEEE Antennas and Propagation Society International Symposium;* 1; (Jul. 13–18); 456–459.

Perrott RA and Griffin JM (1991). L-band antenna systems design. IEE Colloquium on Inmarsat-3; Nov. 21; pp. 5/1–5/7.

Prata Jr A, Filho EA, and Ghosh S (1985). A high performance–wide band–diplexing–tracking–depolarization correcting satellite communication antenna feed. *IEEE MTT-S International Microwave Symposium Digest;* Jun; 477–480.

Rahmat-Samii Y and Galindo-Israel V (1981). Scan performance of dual offset reflector antennas for satellite communications. *Radio Science;* 16; (Nov.–Dec.); 1093–1099.

Rahmat-Samii Y and Imbriale WA (1998). Anomalous results from PO applied to reflector antennas: the importance of near field computations. *Digest, IEEE Antennas and Propagation Society International Symposium;* 2; (June 21–26); 816–819.

Ramanujam P, Lopez LF, Shin C, and Chwalek TJ (1993). A shaped reflector design for the DIRECTV^{TM} direct broadcast satellite for the United States. *Digest, IEEE Antennas and Propagation Society International Symposium;* 2; (June 28–Jul. 2); 788–791.

Ramo S, Whinnery JR, and Van Duzer T (1984). *Fields and Waves in Communication Electronics*, 2nd ed., New York: John Wiley & Sons, Inc.

Rao SK (1999). Design and analysis of multiple-beam reflector antennas. *IEEE Antennas and Propagation Magazine;* 41; (Aug.); 53–59.

Rao SK (2003). Parametric design and analysis of multiple-beam reflector antennas for satellite communications. *IEEE Antennas and Propagation Magazine;* 45; (Aug.); 26–34.

Regier FA (1992). The ACTS multi-beam antenna. *IEEE Transactions on Microwave Theory and Techniques;* 40; 1159–1164.

Roederer A and Sabbadini M (1990). A novel semi-active multi-beam antenna concept. *Digest, IEEE Antennas and Propagation Society International Symposium;* 4; (May 7–11); 1884–1887.

Roper DH, Babiec WE, and Hannan DD (2003). WGS phased arrays support next generation DOD satcom capability. *IEEE International Symposium on Phased Array Systems and Technology;* Oct. 14–17; 82–87.

Rudge AW, Milne K, Olver AD, and Knight P, editors (1982). *The Handbook of Antenna Design*, Vol.1. London:Peter Peregrinus, Ltd.

Rusch WV, Prata Jr A, Rahmat-Samii Y, and Shore RA (1990). Derivation and application of the equivalent paraboloid for classical offset Cassegrain and Gregorian antennas. *IEEE Transactions on Antennas and Propagation;* 38; 1141–1149.

Schennum GH, Space Systems/Loral (2010a). Private communication, Feb. 10.

Schennum GH, Space Systems/Loral (2010b). Private communication, Sept. 23.

Schennum GH, Space Systems/Loral (2011). Private communication, Oct 24.

Schennum GH and Skiver TM (1997). Antenna feed element for low circular cross-polarization. *Proceedings of IEEE Aerospace Conference;* 3; (Feb. 1–8); 135–150.

Schennum GH, Hazelwood JD, Gruner R, and Carpenter E (1999). Global horn antennas for the Intelsat-IX spacecraft. *Proceedings of IEEE Aerospace Conference;* 3; (Mar. 6–13); 221–231.

Schennum GH, Lee E, Pelaca E, and Rosati G (1995). Ku-band spot beam antenna for the Intelsat VIIA spacecraft. *Proceedings of IEEE Aerospace Applications Conference;* 1; (Feb. 4–11); 99–109.

Schuss JJ, Upton J, Myers B, Sikina T, Rohwer A, Makridakas P, Francois R, Wardle L, and Smith R (1999). The Iridium main mission antenna concept. *IEEE Transactions on Antennas and Propagation;* 47; 416–424.

SG Microwave, Inc. (2000). On www.sgmicrowave.com. Accessed Jan. 2009.

Skolnik MI, editor (1970). *Radar Handbook*, New York: McGraw-Hill, Inc.

Sletten CJ and Shore RA (1982). Focal surfaces of offset dual-reflector antennas. *IEE Proceedings, Part H: Microwaves, Optics and Antennas;* 129; (June); 109–115.

Sorrentino R and Bianchi G (2010). *Microwave and RF Engineering*, Chichester, UK: John Wiley & Sons, Ltd.

Stirland SJ and Brain JR (2006). Mobile antenna developments in EADS Astrium. *European Conference on Antennas and Propagation;* Nov. 6–10; pp. 1–5.

Tao ZC, Mahadevan K, Ghosh S, Bergmann J, Sutherland D, and Tjonneland K (1996). Design & evaluation of a shaped reflector & 4-port CP feed for dual band contoured beam satellite antenna applications. *Digest, IEEE Antennas and Propagation Society International Symposium;* 3; (Jul. 21–26); 1692–1695.

Warshowsky J, Kulisan C, and Vail D (2000). 20 GHz phased array antenna for GEO satellite communications. *Proceedings of IEEE Military Communications Conference;* 2; (Oct. 22–25); 1187–1191.

Wertz JR and Larson WJ, editors (1999). *Space Mission Analysis and Design*, 3rd ed., 10th printing 2008. Hawthorne, CA; New York City: Microcosm Press; Springer.

Yodokawa T and Hamada SJ (1981). An X-band single horn autotrack antenna feed system. *IEEE Antennas and Propagation Society International Symposium;* June; 86–89.

Zillig DJ, McOmber DR, and Fox N (1998). TDRSS demand access service: application of advanced technologies to enhance user operations. *SpaceOps Conference;* June 1–5; 1–8.

CHAPTER 4

FILTER AND PAYLOAD-INTEGRATION ELEMENTS

4.1 INTRODUCTION

After discussing the antenna in the previous chapter, we discuss in this chapter the payload filters that are stand-alone units and the passive hardware elements that are used to integrate the payload units together. These are all passive devices. In the following two chapters, we talk about active units, for which nonlinear behavior is a consideration, which does not apply to passive elements; so it makes sense for those two chapters to be consecutive. Also, in this chapter we explain some things that we need in those chapters, for example, insertion loss, return loss, hybrids, and switches.

The only filters we discuss here are the passive filters. Those used in processing satellites, namely, digital and surface-acoustic-wave (SAW), are addressed in Sections 9.3.3 and 9.4.2, respectively.

The rest of the chapter has the following sections:

- *Section 4.2*: *S*-parameters and impedance mismatch
- *Section 4.3*: RF lines for payload integration, namely, coaxial cable and waveguide: their construction, RF performance, environmental considerations, assemblies, and propagation modes
- *Section 4.4*: Other payload integration elements aside from switch
- *Section 4.5*: Filters: types; families of frequency response; general information on resonators and filters; technology with environmental considerations; preselect filter and MUXes; and specification

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

- *Section 4.6*: Switches and redundancy for low-power units and TWTAs (for SSPAs in Section 6.5.1.2)
- *Appendix 4.A*: Filter poles and zeros.

4.2 IMPEDANCE MISMATCH

Impedance mismatch is a topic that must be mastered for proper accounting of the losses accrued during payload integration, which are not assignable to the units and integration elements themselves but to the fact that they are connected with each other. When two elements are connected, power is transferred from the first element (the "source") to the other (the "load") without loss if and only if the load impedance equals the complex conjugate of the source impedance. However, the match is never perfect, so some power reflects backward. Units are designed with the intention of making a good match with adjacent elements; one specification common to all elements is input return loss (defined below), to limit the reflected power.

Let us approach the topic by starting with **scattering parameters (*S*-parameters)**, which may be known to the payload engineer, and go on to impedance mismatch. Suppose that a traveling voltage-wave with complex phasor (Section 10.2.1.2) V_1^+ is incident at port 1 of a two-port "device under test" (DUT). A voltage V_2^- is transmitted from port 2, and a voltage V_1^- is reflected out of port 1. The output line is terminated in such a manner that there is no incident voltage on port 2. This situation is depicted in Figure 4.1a. Then two of the complex-valued scattering parameters are defined from this, as follows:

$$S_{21} \triangleq \frac{V_2^-}{V_1^+} = \text{Forward transmission coefficient}$$

$$S_{11} \triangleq \frac{V_1^-}{V_1^+} = \text{Input-port reflection coefficient}$$

Similarly, suppose that the incident traveling wave is incident on port 2 and that port 1 is terminated so that there is no incident voltage on it. This situation is depicted in Figure 4.1b. Then the other two complex-valued scattering parameters

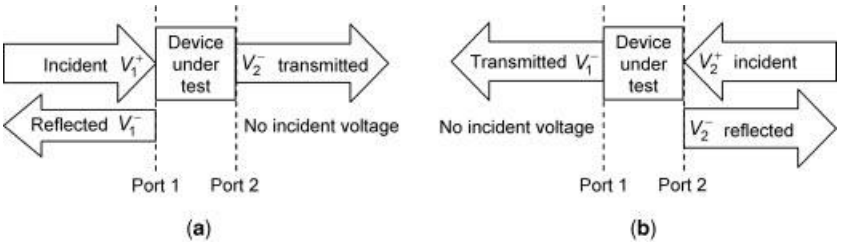


FIGURE 4.1 Voltages in definition of *S*-parameters: (a) for S_{21} and S_{11} and (b) for S_{12} and S_{22}.

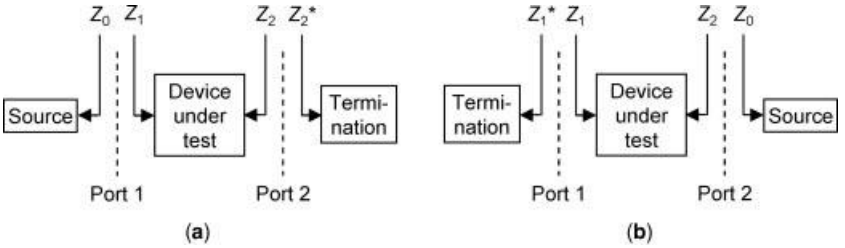


FIGURE 4.2 Conceptual impedance setups for obtaining *S*-parameters: (a) for S_{21} and S_{11} and (b) for S_{12} and S_{22}.

are defined from this, as follows:

$$S_{12} \triangleq \frac{V_1^-}{V_2^+} = \text{Reverse transmission coefficient}$$

$$S_{22} \triangleq \frac{V_2^-}{V_2^+} = \text{Output-port reflection coefficient}$$

The short-hand definition of S_{mn} is that it is the output of port m from input on port n, relative to the input. The impedance setups for measuring the *S*-parameters are shown in Figure 4.2. The device under test has impedance Z_1 at input and Z_2 at output. S_{21} and S_{11} are obtained using a source at a reference impedance Z_0 and an ideal output termination, that is, with impedance Z_2^* equal to the complex conjugate of Z_2. S_{12} and S_{22} are obtained in a complementary fashion. For coaxial input, the reference impedance is typically $50\,\Omega$.

The complex value of the *S*-parameter S_{mn} is usually stated as a gain $|S_{mn}|$ and a phase. The gain $|S_{mn}|$, not in dB, may be any positive number. In particular, $|S_{21}|$ is the **insertion gain**. For passive devices $|S_{21}|<1$, actually representing a loss; the term then is **insertion loss**, equal to $1/|S_{21}|$, which is greater than 1. Normally, $|S_{11}|$ and $|S_{22}|$ are less than 1. Their inverses are respectively the **input and output return losses**. For all these gain or loss terms, their normal representation is not as a simple ratio but in dB, namely, for example, $-20\log_{10}|S_{21}|$ for the insertion loss of a passive device.

We can now express impedance mismatch in terms of the *S*-parameters. The complex-valued "voltage reflection coefficient" Γ_1 at port 1 equals S_{11}. There is a standing wave on the input RF line to the DUT, consisting of the complex sum of the incident and reflected waves (see, e.g., Ramo et al. (1984)). The **voltage standing-wave ratio** (**VSWR**) at port 1 is the ratio of the maximum voltage amplitude $\left(\left|V_1^+\right| + \left|V_1^-\right|\right)$ on the RF line to the minimum voltage amplitude $\left(\left|V_1^+\right| - \left|V_1^-\right|\right)$. It is not given in dB; sometimes, to reinforce that point, it is stated in the format "x:1" where x is the VSWR. Thus,

$$\text{VSWR at port 1} = \frac{1+|\Gamma_1|}{1-|\Gamma_1|}, \qquad |\Gamma_1| = \frac{(\text{VSWR at port 1})-1}{(\text{VSWR at port 1})+1}$$

TABLE 4.1 VSWR with Corresponding Return Loss and Mismatch Loss

VSWR	Return Loss (dB)	Mismatch Loss (dB)	VSWR	Return Loss (dB)	Mismatch Loss (dB)
1.00	∞	0	1.24	19.40	0.050
1.02	40.09	0.0004	1.26	18.78	0.058
1.04	34.15	0.0017	1.28	18.22	0.066
1.06	30.71	0.0037	1.30	17.69	0.075
1.08	28.30	0.0064	1.35	16.54	0.097
1.10	26.44	0.0099	1.40	15.56	0.122
1.12	24.94	0.014	1.45	14.72	0.149
1.14	23.69	0.019	1.50	13.98	0.177
1.16	22.61	0.024	1.60	12.74	0.238
1.18	21.66	0.030	1.70	11.73	0.302
1.20	20.83	0.036	1.80	10.88	0.370
1.22	20.08	0.043	1.90	10.16	0.430

Similarly, the VSWR Γ_2 at port 2 equals S_{22}. **Mismatch loss** is the ratio of power transmitted to power incident in the forward direction, namely $1/(1-|S_{11}|^2)$ or equivalently $1/(1-|\Gamma_1|^2)$. It is typically stated in dB, equaling $-10\log_{10}(1-|\Gamma_1|^2)$. Conceptually every connection causes a mismatch loss. Table 4.1 gives a list of VSWR values each with corresponding return loss and mismatch loss.

If there are two discontinuities in a signal path, a wave reflected from the second discontinuity is then reflected forward at the first discontinuity. This creates "multipath" because two versions of the signal continue past the second discontinuity, with the doubly reflected one being delayed and weaker. A broken piece of waveguide can be the culprit. It is often easy to find which piece of waveguide is responsible. The $|S_{21}|$ response versus frequency will be periodic, looking more or less like a sinewave, whose frequency period gives away the length of the responsible waveguide piece (Appendix 7.A.2).

4.3 RF LINES FOR PAYLOAD INTEGRATION

The types of RF line we discuss in this section are coaxial cable (coax) and waveguide. These are the only kinds of line used to integrate the payload. There is no agreement on a term that includes both coaxial cable and waveguide. In some professional books and articles, they are both called "transmission lines" or "transmission structures." In textbooks, coaxial cable is a "transmission line" when it is propagating TEM mode (Section 4.3.2.5), and both coaxial cable and waveguide are "waveguides."

When to use coax and when to use waveguide is usually clear. Waveguide has higher mass per meter of run length than coax does for frequencies below about 10 GHz, while above about 10 GHz the masses are roughly the same, as will be

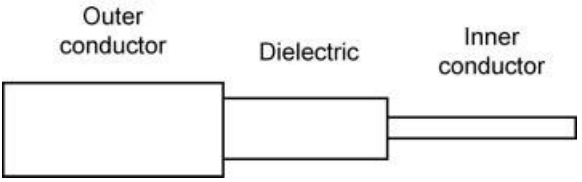


FIGURE 4.3 Coaxial cable electrical construction.

seen in the tables discussed below. Waveguide is larger in size than coax at all frequencies, and the lower the frequency the greater the difference. The insertion loss per meter of waveguide is much lower than that of coax. The power-carrying ability of waveguide is higher. The length tolerance of coax can be greater than for waveguide, since coax can be bent a little but waveguide cannot. In general, coax is used at low frequencies and low power, while waveguide is used at the higher frequencies or when the power is high.

4.3.1 Coaxial Cable

4.3.1.1 Coax Construction Coax consists, electrically, of three elements, as shown in Figure 4.3, a center conductor, an outer conductor, and a dielectric filling the space in between. The dielectric has the main purpose of altering the electrical properties but it also serves the mechanical purpose of keeping the conductors properly separated. The inner conductor is usually silver-plated copper wire. At GHz frequencies, only the outer surface of the inner conductor and the inner surface of the outer conductor conduct. The outer conductor is made of copper or aluminum alloy. Coating or finish of the outer conductor is on the outside and does not affect the electrical performance. Semirigid coax is the kind used for payload integration except for applications such as a gimballed antenna, which call for flexible (WL Gore, 2003). We do not discuss flexible coax further.

The dielectric is always some form of polytetrafluoroethylene (PTFE), known otherwise by the DuPont brand name Teflon (Wikipedia, 2011c). There are three forms of it in use in coax:

- *Solid*: The oldest form. Lowest in mass.
- *Low-Loss Low-Density*: Intermediate in mass.
- *Expanded or Microporous, Sometimes Written "ePTFE"*: Least rugged (Teledyne, 2010). Greatest in mass.

Tables 4.2–4.4 provide examples of coax structure for the three dielectrics in order, in various sizes. In all cases, the center conductor is silver-plated copper wire. One column in each table gives the nominal mass of a coax run in kg per 10 m, for comparison purposes with other coax and with waveguide. Coax with the aluminum-alloy outer conductor is lower in mass than with copper.

TABLE 4.2 Examples of Coax Dimension and Mass, Solid PTFE

Outer-Conductor Inner Diameter (in.)	Max Operating Freq. (GHz)	Outer Conductor Material	Nominal Mass (kg/10 m)
0.215	<21	OFHC[a] Cu	1.024
0.141	<34	OFHC Cu; Al alloy	0.467; 0.308
0.085	<61	OFHC Cu; Al alloy	0.202; 0.113
0.047	<108	OFHC Cu; Al alloy	0.057; 0.032

Source: Micro-Coax (2008).
[a] Oxygen-free high thermal-conductivity.

TABLE 4.3 Examples of Coax Dimension and Mass, Low-Loss Low-Density PTFE

Outer-Conductor Inner Diameter (in.)	Max Operating Freq. (GHz)	Outer Conductor Material	Nominal Mass (kg/10 m)
0.141[a]	26.5	OFHC Cu	0.486
0.086[a]	40 or 65 (two versions)	OFHC Cu	0.228
0.070[b]	<72	Cu	0.115

[a] Teledyne (2010).
[b] Micro-Coax (2008).

The manufacturer preconditions coax by heating it. After the coax is bent for payload integration and before installation, it may need to be conditioned again by three thermal cycles, where in each cycle the temperature goes from maximum to ambient to minimum to ambient. The conditioning will make the dielectric retract, so the cable should originally be cut $\frac{1}{4}$ in. longer than its design length. After conditioning, the cable is cut to its final length and the connectors put on (Micro-Coax, 2008).

TABLE 4.4 Examples of Coax Dimension and Mass, Expanded PTFE

Outer-Conductor Inner Diameter (in.)	Max Operating Freq. (GHz)	Outer Conductor Material	Nominal Mass (kg/10 m)
0.320	18	Cu	1.444
0.290	18	Cu	1.312
0.210	18 (phase-stable)	Cu	0.627
0.190	18 or 30 (two options)	Cu	0.558
0.140	40	Cu	0.328
0.120	26.5	Cu	0.295
0.085	65	Cu	0.131
0.047	65	Cu	0.065

Source: WL Gore (2003).

4.3.1.2 Coax Performance

According to the dielectric material, the RF performance is summarized as follows:

- *Solid*: Highest loss.
- *Low-Loss Low-Density*: Intermediate in loss. Exhibits greater variation in dielectric constant over length and among lots than solid does.
- *Expanded or Microporous*: Least lossy. Exhibits greatest variation in dielectric constant over length and among lots (Teledyne, 2010).

Cable comes in a family of sizes, each of which can be used from DC to some upper frequency. The higher the upper frequency is, the higher the cable's insertion loss. Smaller-size coax is not necessarily made for higher frequencies, as is true for waveguide, but for lower mass; it has higher insertion loss. For a given piece of coax, the insertion loss increases with frequency in a nonlinear fashion (Teledyne, 2010) and the insertion phase increases linearly. Table 4.5 gives some examples of coax insertion loss at ambient. The typical impedance of coax is 50 Ω but lower- and higher-impedance coax is available.

4.3.1.3 Coax Environmental

The temperature sensitivity of coax depends on the dielectric material and the coax size (Teledyne, 2010). The temperature performance of the three dielectrics is as follows:

- *Solid*: Most temperature-sensitive.
- *Low-Loss Low-Density*: Intermediate in temperature-sensitivity.

TABLE 4.5 Examples of Typical Coax Insertion Loss at Ambient

Outer-Conductor Inner Diameter (in.)	Dielectric	Outer Conductor Material	At 1 GHz, Insertion Loss per 10 m (dB)	At 10 GHz, Insertion Loss per 10 m (dB)	At 20 GHz, Insertion Loss per 10 m (dB)
0.140[a]	Expanded PTFE	Cu	4	12	18
0.141[b]	Low-loss low-density PTFE	Cu	3	11	16
0.141[c]	Solid PTFE	Cu	4	14	21
0.141[c]	Solid PTFE	Al alloy	4	14	22
0.086[b]	Low-loss low-density PTFE	Cu	5	20	29
0.070[c]	Low-loss low-density PTFE	Cu	7	21	30
0.047[c,d]	Low-loss low-density PTFE	Cu	10	33	47

[a] WL Gore (2003).
[b] Teledyne (2010).
[c] Micro-Coax (2008).
[d] Center conductor Cu but not wire, but irrelevant to insertion loss.

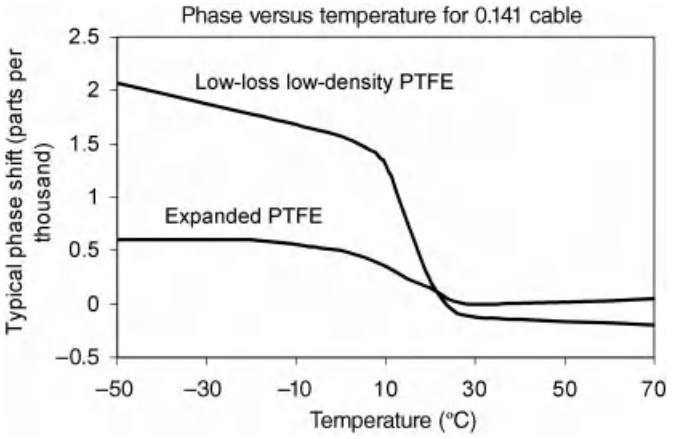


FIGURE 4.4 Example of coax electrical length variation versus temperature (after Teledyne (2010)).

- *Expanded or Microporous*: Least temperature-sensitive. Most phase-stable, is used when a set of coax must track in phase and gain over temperature.

As temperature goes up, coaxial cable's insertion loss increases. One cable manufacturer reports that the percentage change in dB loss varies nearly linearly with temperature, relative to the dB loss at 25°C. The variation is −22% at −100°C and 18% at 150°C, for 18 GHz. The variation is a weak function of frequency. The primary cause is the change in conductivity of silver with temperature (WL Gore, 2003).

In addition, the insertion phase changes with temperature. An example of this is shown in Figure 4.4. The insertion phase also changes with handling–bending the cable, even a little bit by removing it and reinstalling it, will cause a phase change (WL Gore, 2003).

A set of expanded-PTFE cable can be either relatively phase-matched or absolutely phase-matched. Being able to phase-match is an important consideration when the cables must track over temperature in regard to insertion loss and/or phase. The cables have to have the same length and be at the same temperature. The tightest tracking may be achieved when the coax is all from the same material batch (WL Gore, 2003).

In space, there will be a brief period of time during which the dielectric vents. This is called "out-gassing." Spaceflight coax meets standards on the amount of outgassing (Micro-Coax, 2001; WL Gore, 2003).

Spaceflight coax is designed to be radiation-tolerant (Micro-Coax, 2001; WL Gore, 2003).

4.3.1.4 Connector and Adapter

There are many kinds of **connector** available to put on the ends of coax runs. The choice depends partly on the coax size and

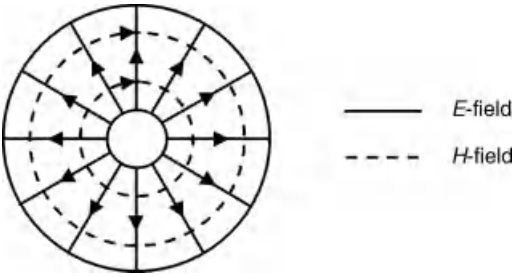


FIGURE 4.5 TEM propagation mode in coax (after Sorrentino and Bianchi (2010)).

the operating frequency range. Even then, each type has several configurations, for example, straight or at a right angle. Care must be taken to use connectors with the appropriate recommended operating frequency range. Some connectors of different types can mate with each other. Connectors also go on payload units. Connectors are described in Teledyne (2010) and WL Gore (2003).

An **adapter** converts between waveguide with a given flange and coax with a given connector. It converts between the propagation modes.

4.3.1.5 Propagation in Coax Only the transverse electromagnetic (TEM) mode propagates in coax. The electric field vector $\boldsymbol{E}$ and its orthogonal magnetic field vector $\boldsymbol{H}$ lie in the wavefront plane, that is, are perpendicular to the direction of propagation, and the fields' amplitudes are related by a constant (Ramo et al., 1984). The fields are shown in Figure 4.5. Recall that TEM is also the mode in which radiation propagates in space (Section 3.2.3.2). Higher-order modes than TEM could exist in coax, but the ratio of the conductor diameters is chosen so that the cutoff frequencies of the higher-order modes are well above the operating frequencies (Ramo et al., 1984).

Coax is a nondispersive medium, that is, all frequencies propagate at the same speed. Light in coax has a **velocity of propagation** that is slower than in vacuum. It equals the speed of light c in vacuum divided by the square root of the dielectric constant. This is about 70% of c for solid PTFE, 75–79% for low-loss low-density PTFE, and 82–88% for expanded PTFE, as shown in Table 4.6 (Teledyne, 2010). As a general rule, the lower the dielectric constant, the lower the loss (Micro-Coax, 2008).

TABLE 4.6 Velocity of Propagation in Coax with Various Dielectrics

Coax Dielectric	Dielectric Constant	Velocity of Propagation (% of c)
Solid PTFE	2.02	70
Low-loss low-density PTFE	1.6–1.8	75–79
Expanded PTFE	1.3–1.5	82–88

Source: Teledyne (2010).

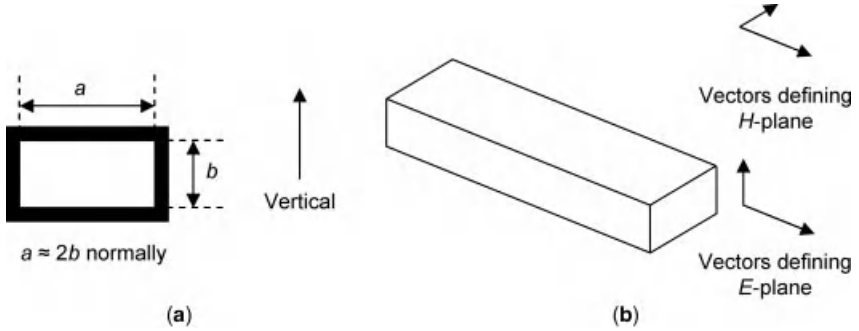


FIGURE 4.6 Rectangular waveguide (a) dimensions and (b) *E*- and *H*-planes.

4.3.2 Waveguide

Rectangular, not circular, waveguide is most often used in integration. Circular waveguide is, however, of interest because its propagation modes are the basis of the resonance modes in some kinds of filter (Section 4.5). Besides, some antenna horns (Section 3.4.1) are flared circular waveguide.

4.3.2.1 Rectangular Waveguide Construction Some defining characteristics of rectangular waveguide are shown in Figure 4.6. The waveguide's inner dimensions are width a and height b, where in normally proportioned waveguide a is about equal to twice b. The second part of the figure shows the so-called *E*- and *H*-planes.

Table 4.7 gives the specifications of rectangular waveguide in its various standard sizes. The size designation is "WR" followed by the width a in hundreds of inches. Most of the waveguides have the normal or nominal dimensions. However, those with "R/H" at the end of their designation have reduced height, about half the normal height. Reduced-height waveguide has a little lower mass than normal waveguide, and its small volume may be a benefit in a tight fit. There are also quarter-height waveguide and double-height waveguide (not listed). There is a normal or nominal wall-thickness. However, thin-wall waveguide, with half the normal thickness, is typically used on spacecraft for the reduced mass (Mele, 2011). The thin-wall thickness is shown in Table 4.7 for some waveguide sizes. Aluminum is the material commonly used on spacecraft instead of copper, when possible, because aluminum has about one-third the mass of copper. In Table 4.7, the smallest waveguides are not available in aluminum, perhaps because of the very high insertion loss. The recommended operating band of frequencies depends on the dimensions a and b. So does the theoretical insertion loss, which also depends on the metal and the exact operating frequency. For a given piece of waveguide, the insertion loss decreases with increasing frequency.

Waveguide can also be made of graphite fiber-epoxy composite (GFEC), consisting of layers of graphite fiber bonded with interleaved layers of cured epoxy resin.

TABLE 4.7 Rectangular Waveguide Specifications, Thin-Wall

EIA WG Designation	Recommended Operating Band for TE_{10} Mode (GHz)	Material Alloy	Theoretical Insertion Loss, Lowest to Highest Frequency (dB/10 m)	*a* (in.)	*b* (in.)	Thin-Wall Thickness (in.)	Mass for Al Alloy (kg/10 m)
WR430	1.70–2.60	Al; Cu	0.194–0.129; 0.129–0.086	4.300	2.150	–	–
WR340	2.20–3.30	Al; Cu	0.263–0.183; 0.175–0.122	3.400	1.700	–	–
WR284 R/H	2.60–3.95	Al; Cu	0.585–0.442; 0.389–0.294	2.840	0.670	–	–
WR284	2.60–3.95	Al; Cu	0.366–0.251; 0.243–0.167	2.840	1.340	–	–
WR229	3.30–4.90	Al; Cu	0.467–0.331; 0.310–0.220	2.290	1.145	–	–
WR187	3.95–5.85	Al; Cu	0.688–0.477; 0.458–0.317	1.872	0.872	–	–
WR159	4.90–7.05	Al; Cu	0.766–0.572; 0.503–0.381	1.590	0.795	–	–
WR137	5.85–8.20	Al; Cu	0.969–0.770; 0.652–0.512	1.372	0.622	–	–
WR112	7.05–10.00	Al; Cu	1.369–1.062; 0.911–0.707	1.122	0.497	0.040/0.045	2.54
WR102	7.00–11.00	Al; Cu	1.734–1.094; 1.154–0.727	1.020	0.510	–	–
WR96	7.00–17.00	Al; Cu	2.878–1.701; 1.914–1.130	0.965	0.320	–	–
WR90 R/H	8.20–12.40	Al	3.435–3.036	0.900	0.200	–	–
WR90	8.20–12.40	Al; Cu	2.135–1.477; 1.390–0.983	0.900	0.400	0.030/0.035	1.56
WR75 R/H	10.00–15.00	Al	2.561–1.952	0.750	0.200	–	–
WR75	10.00–15.00	Al; Cu	2.526–1.764; 1.680–1.174	0.750	0.375	0.020/0.025	0.93
WR67	11.00–17.00	Al; Cu	2.029–1.354; 1.349–0.900	0.668	0.340	–	–
WR62	12.40–18.00	Al; Cu	3.182–2.340; 2.116–1.556	0.622	0.311	0.020/0.025	0.78
WR51	15.00–22.00	Al; Cu	4.347–3.149; 2.891–2.094	0.510	0.255	–	–
WR42	18.00–26.50	Al; Cu	6.804–4.997; 4.528–3.323	0.420	0.170	0.020/0.025	0.51
WR34	22.00–33.00	Al; Cu	8.317–5.784; 5.531–3.848	0.340	0.170	–	–
WR28	26.50–40.00	Al; Cu	11.306–7.740; 7.552–5.174	0.280	0.140	0.020/0.025	0.40
WR22	33.00–50.00	Cu	10.643–7.234	0.224	0.112	–	–
WR19	40.00–60.00	Cu	13.061–9.383	0.188	0.094	–	–

Source: After Cobham (2006).

TABLE 4.8 Example of Rectangular Waveguide Insertion-Loss Specification, Worst Over Recommended Operating Band

EIA WG Designation	Rigid-Aluminum Insertion Loss (dB/10 m)	Rigid-OFHC Copper[a] Insertion Loss (dB/10 m)
WR430	0.4	–
WR340	0.4	–
WR284 R/H	0.8	–
WR284	0.4	0.4
WR229	0.4	0.4
WR187	0.8	0.4
WR159	0.8	0.8
WR137	1.2	0.8
WR112	1.6	1.2
WR102	1.6	1.2
WR96	3	2.0
WR90 R/H	4	–
WR90	2.0	1.6
WR75 R/H	5	–
WR75	2.4	2.0
WR67	3	2.4
WR62	3	2.4
WR51	5	4
WR42	8	6
WR34	12	6
WR28	16	8
WR22	–	12
WR19	–	12

Source: After Cobham (2006).
[a] Oxygen-free high thermal-conductivity copper.

The waveguide receives a silver-plating (Kudsia and O'Donovan, 1974). Its mass is much lower than that of aluminum waveguide. It was used on the Sirius Radio satellites at *S*-band to combine the outputs of 32 TWTAs in phase (Briskman and Prevaux, 2001). Its disadvantages are that it is expensive and fragile and it cannot be dent-tuned (Section 4.3.2.4).

4.3.2.2 Rectangular Waveguide Performance Table 4.8 gives the insertion-loss specification of one manufacturer for the various sizes of rectangular waveguide (Cobham, 2006). The loss is the same for both wall thicknesses. Reduced-height waveguide has about double the dB insertion loss of normal-height waveguide. The specification is for aluminum and oxygen-free high thermal-conductivity (OFHC) copper. Aluminum's insertion loss is about 1.5 times that of copper. Silver-plating aluminum is a way to reduce the insertion loss close to that of copper. However, silver-plating is not possible in the smallest waveguide sizes. The specified insertion losses range from between one and two times the theoretical insertion loss. Waveguide insertion loss varies more over a frequency band than

coax insertion loss does. The dB insertion loss of rectangular waveguide decreases nonlinearly with frequency over its recommended operating band, and the insertion phase is nonlinear (Ramo et al., 1984).

4.3.2.3 Waveguide Environmental An aluminum or copper waveguide run or "tube" expands as the waveguide material does, with temperature. The frequency response shifts to lower frequencies. In situations where this is not acceptable, for example, where multiple HPA outputs must track in phase, graphite waveguide can be used. The graphite layers have a positive CTE, while the epoxy layers have a negative CTE. The waveguide's coefficient of thermal expansion (CTE) is like that of the filter material Invar (Section 4.5.2.1) (Kudsia and O'Donovan, 1974), namely almost as small as one-twentieth (NIST, 2004) of the CTE of the aluminum alloy used by one manufacturer for thin-wall aluminum waveguide (Cobham, 2006).

4.3.2.4 Flange and Waveguide Assemblies Waveguide runs receive **flanges** on their ends, which correspond to coax connectors. Flanges come in many sizes and shapes. Some payload units also have flanges.

In order to be able to connect waveguide runs or units to waveguide runs, other waveguide elements are needed such as the following:

Bend, in either the *E*-plane or the *H*-plane

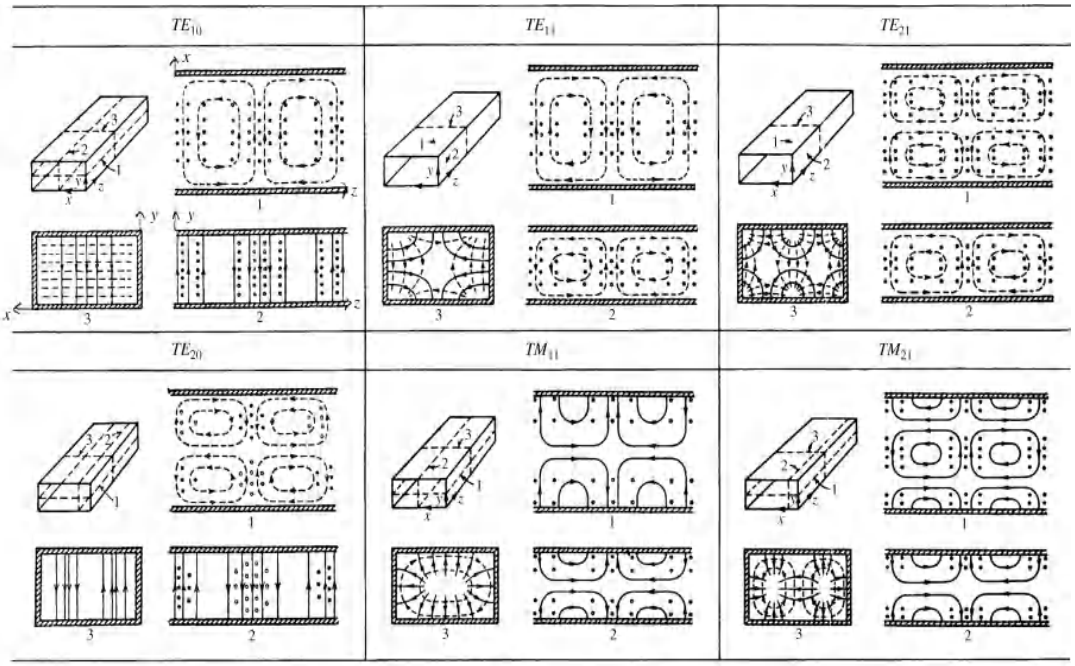
Transition, which are used to connect two different sizes of waveguide

Twist, to change the field orientation.

Dent-tuning is sometimes performed on a waveguide assembly to improve its VSWR. A dent is hammered into the appropriate place by a very experienced person. Graphite waveguide cannot be dent-tuned.

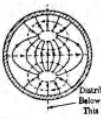
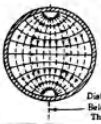
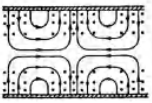
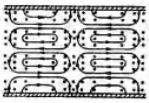
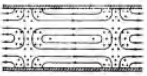
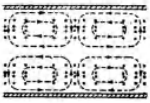
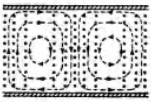
4.3.2.5 Propagation in Rectangular and Circular Waveguide The two forms of waves or modes that propagate in waveguide are transverse electric (TE) and transverse magnetic (TM). In TE the electric field is transverse to the direction of propagation, while in TM the magnetic field is transverse to the propagation direction. There is a family of TE modes and a family of TM modes, both different for rectangular and circular waveguides. The modes have the same notation for both shapes, though, namely TE_{mn} and TM_{mn}. For each index pair (m,n) there is a **cutoff frequency**, dependent on the waveguide dimension a and the dielectric filling the waveguide, below which there is no propagation. The dielectric is air on the ground and vacuum in space. The dominant mode is the one with the lowest operating frequency.

Specifically, for rectangular waveguide of normal proportions (Figure 4.6), the TE_{mn} mode has m half-wave variations of the electric field across a and n half-wave variations across b. There is a similar family TM_{mn}. The dominant mode is TE_{10}. For circular waveguide, the TE_{mn} mode has m variations of the electric field circumferentially and n radially. There is a similar family of TM modes, TM_{mn}. The dominant mode is TE_{11}. The first few modes for rectangular waveguide of normal proportions are shown in Figure 4.7 and for circular waveguide in Figure 4.8.



[a] Electric field lines are shown solid and magnetic field lines are dashed.

FIGURE 4.7 Propagation modes in rectangular waveguide. (Reprinted with permission of John Wiley & Sons, Inc., from Ramo et al. (1984)).

Wave Type	TM_{01}	TM_{02}	TM_{11}	TE_{01}	TE_{11}
Field distributions in cross-sectional plane, at plane of maximum transverse fields					
Field distributions along guide					
Field components present	E_z, E_r, H_ϕ	E_z, E_r, H_ϕ	$E_z, E_r, E_\phi, H_r, H_\phi$	H_z, H_r, E_ϕ	$H_z, H_r, H_\phi, E_r, E_\phi$
p_{nl} or p'_{nl}	2.405	5.52	3.83	3.83	1.84
$(k_c)_{nl}$	$\frac{2.405}{a}$	$\frac{5.52}{a}$	$\frac{3.83}{a}$	$\frac{3.83}{a}$	$\frac{1.84}{a}$
$(\lambda_c)_{nl}$	$2.61a$	$1.14a$	$1.64a$	$1.64a$	$3.41a$
$(f_c)_{nl}$	$\frac{0.383}{a\sqrt{\mu\varepsilon}}$	$\frac{0.877}{a\sqrt{\mu\varepsilon}}$	$\frac{0.609}{a\sqrt{\mu\varepsilon}}$	$\frac{0.609}{a\sqrt{\mu\varepsilon}}$	$\frac{0.293}{a\sqrt{\mu\varepsilon}}$
Attenuation due to imperfect conductors	$\frac{R_s}{a\eta}\frac{1}{\sqrt{1-(f_c/f)^2}}$	$\frac{R_s}{a\eta}\frac{1}{\sqrt{1-(f_c/f)^2}}$	$\frac{R_s}{a\eta}\frac{1}{\sqrt{1-(f_c/f)^2}}$	$\frac{R_s}{a\eta}\frac{(f_c/f)^2}{\sqrt{1-(f_c/f)^2}}$	$\frac{R_s}{a\eta}\frac{1}{\sqrt{1-(f_c/f)^2}}\left[\left(\frac{f_c}{f}\right)^2+0.420\right]$

[a] Electric field lines are shown solid and magnetic field lines are dashed.

FIGURE 4.8 Propagation modes in circular waveguide. (Reprinted with permission of John Wiley & Sons, Inc., from Ramo et al. (1984)).

The fact that waveguide has a cutoff frequency is used to ensure that there is no ring-around of the payload's transmit signal into the payload's receiver. The receive antenna's terminal is made, wholly or in part, of waveguide selected in size to cut off the transmit band, which is lower than the receive band (Schennum, 2011).

A hybrid mode is a combination of a TE mode and a TM mode. In circular waveguide, the dominant hybrid mode HE_{11} is the combination of TE_{11} and TM_{11}. A **degenerate mode** is one of a set of propagation modes that have the same phase velocity but different field configurations in any transverse cross-section.

The **group velocity** of the propagating wave is the speed at which power progresses down the waveguide (Chang, 1989). For each waveguide size, the group velocity varies from roughly $0.6c$ at the low end of the recommended operating frequency band to about $0.85c$ at the high end, where c is the speed of light in vacuum (Cobham, 2006). Thus, group velocity is dependent on frequency; waveguide causes **dispersion**. The **phase velocity** is the speed at which a constant phase of a wave travels down the waveguide. It is greater than c, and the product of phase velocity and group velocity is a constant independent of frequency. See Section 8.4.3 for a way of approximately phase-compensating a length of waveguide with a length of coax.

For further information, see, for example, Chang (1989) or Ramo et al. (1984).

4.4 OTHER PAYLOAD-INTEGRATION ELEMENTS ASIDE FROM SWITCH

The payload-integration items addressed here are used in close conjunction with filters and are sometimes part of filter assemblies.

The unexciting but crucial **termination** is a load designed to absorb the power sent to it and not reflect it back. Its impedance must be matched to whatever it is terminating. A termination must be designed to absorb the highest power that could be sent to it. There are low- and high-power terminations, of coax and waveguide. We will see that terminations are used with hybrids and switches to terminate the unused ports; the purpose is to improve the performance of the other ports.

A **circulator** is a device with three ports that sends a signal input on one port out on the next port in circular order, as shown in Figure 4.9. Circulators are used in most input multiplexers. The circulator must be able to accommodate the RF power that it could see. Circulators come in both coax and waveguide versions. The coax versions are either designed for an extremely wide range of frequencies, with VSWR that increases with frequency, or over a lesser range but with better VSWR on that range. A waveguide circulators' frequency range is restricted by its flange, that mates with a particular size of waveguide.

An **isolator** is equivalent, RF-circuit-wise, to a circulator with a termination on one port. However, an isolator and a circulator-with-termination have slightly different implementations that can matter in a high-power application: the termination that is part of an isolator is right with the rest of the isolator, while the termination on a circulator can be some distance away from the circulator, making it easier to

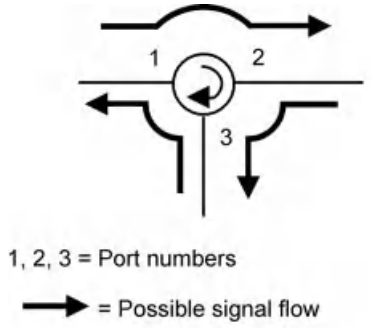


FIGURE 4.9 Circulator.

site the termination on a heat pipe. An isolator is used to improve the impedance match of a unit or to protect a unit from backwards reflections into it. Figure 4.10a and b show examples of an isolator used for impedance matching. Figure 4.10c shows an isolator placed so as to protect a HPA from a backwards-propagating signal that could result from a mis-switch past the HPA.

A hybrid coupler or simply **hybrid** is a device that combines two inputs, yielding two outputs, where each output is half the sum of both inputs in some particular phase relationship. A **power divider** divides one input into two equal ones. It can be used in reverse so that instead two inputs are combined into one output, making it a **power combiner**. Even though their implementations may be different, a power divider is conceptually the same as a hybrid with one port terminated. Usually in this book, we use a hybrid symbol "H" to mean either a hybrid or a power divider. A hybrid is a 3-dB coupler. A hybrid's or power divider's insertion loss specification is usually given relative to the unavoidable 3-dB loss. The hybrid must be able to carry the right amount of power.

There are various kinds of hybrid. There is an overlap between the kinds implemented in waveguide and the kinds implemented in coax. The difference among hybrids is the difference between input and output phases. Figure 4.11 shows the phase relationships for one kind of hybrid. For another kind, see Section 6.4.2.1. By means of adding 90° phase-shifters, one can make all of them do the same thing. When only one signal, say x in Figure 4.11, is input to the hybrid, then half the power of the signal goes to each output port. When the two input signals x and y are identical and in phase, then the hybrid has only one nonzero output, equal to x with double power.

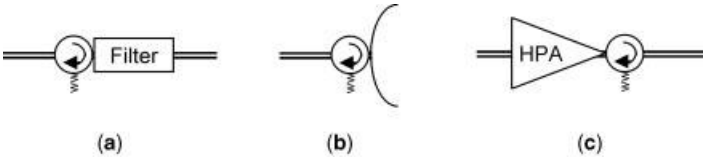


FIGURE 4.10 Isolator usage examples.

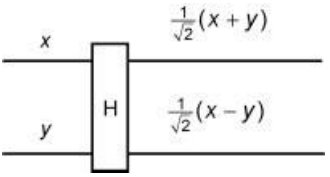


FIGURE 4.11 Lossless hybrid, showing input and output RF signals.

Fixed-value attenuators, also called **pad attenuators**, are coaxial. Each is usable from DC to some GHz frequency, for example, for one manufacturer to either 18 or 32 GHz (Aeroflex, 2010). Available attenuation values are from 0 dB in 0.5-dB steps. These attenuators can carry 2 W. Fixed-value attenuators are used in payload integration to align the operating signal-level ranges of successive units. They also have the purpose of improving impedance match. Such a use can be seen in Figure 4.12. There are units A and B separated by an attenuator, say a 3-dB attenuator. The intended signal, signal 1, has a level at input to unit B that is 3 dB lower than it would be without the attenuator. If there is a reflected signal from unit B back to unit A, the signal arriving at unit A's output port is 6 dB lower than it would be without the attenuator, so it is less harmful to unit A's performance. Furthermore, if there is a doubly reflected signal, signal 3, when it arrives at the input to unit B it is 9 dB lower than it would be without the attenuator, namely 6 dB lower than the intended signal. The attenuator could as well be at input to unit B instead of at output of unit A.

Phase-shifters are coaxial and are adjustable during payload integration. One example of a phase-shifter is designed for DC to 18 GHz and has a phase shift range of 10 times the carrier frequency in GHz, namely, less than 360° of tunability (Radiall, 2011b). Phase-shifters are needed to put TWTA outputs that will be combined, in phase with each other at the combiner. The method for this is as follows. The TWTA manufacturer measures the phase shifts of the TWTAs. The payload integrator measures the phase shift in the waveguide from a TWTA to the post-TWTA combiner. The calculated target phase shift for the coax run from the (L)CAMP to TWTA input, including the phase shifter, is then computed and the phase shifter is accordingly adjusted by means of a vector network analyzer.

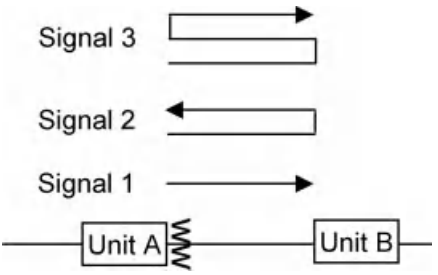


FIGURE 4.12 Fixed attenuator improving impedance matches.

4.5 FILTER

This section on filter units is a large section because filters are complicated. We start off with a listing of the various types of filter, defined by their passband. This includes one kind probably unknown to most readers, the all-pass filter. The next subsection is about the most common filter responses used in the payload, the Chebyshev, elliptic, and pseudoelliptic. Readers curious about poles and zeros and the Laplace transfer function of a filter may read Appendix 10 at that point. Section 4.5.1 gives general information about resonators, filters, and multiplexers. Section 4.5.2 is about the three filter technologies used most in the frequency regime of 1 GHz and up, namely the empty-cavity waveguide filter, the dielectric-resonator loaded-cavity waveguide filter, and the coaxial-cavity combline filter. Section 4.5.3 is about the specifics of filters in the payload. Section 4.5.4 lists the communications parameters one should expect in a specification.

Payload engineers usually think about a filter in terms of its (Fourier) **transfer function** $H(f)$ in the frequency domain, where f is frequency in Hz (for more details see Chapter 10). The transfer function is in general complex-valued. It consists of two parts, the **gain response** $G(f)$ and the **phase response** $\varphi(f)$:

$$H(f) = 10^{G(f)/20} e^{j\varphi(f)}$$

where $G(f)$ is in dB and $\varphi(f)$ is in radians.

However, filter engineers think about a filter in terms of its Laplace transfer function, which is a function of the complex variable $s = \sigma + j\omega = \sigma + j2\pi f$ where σ is real. The Fourier transfer function equals the Laplace transfer function on the $j\omega$-axis.

4.5.1 General

4.5.1.1 Filter Types The five types of filter are the following, depending on what contiguous band of frequencies the filter passes:

- ***Low-Pass Filter*** **(LPF)**, which passes a contiguous band of frequencies including DC
- ***High-Pass Filter*** **(HPF)**, which passes all frequencies above some particular frequency. These are uncommon except that waveguide itself is a HPF
- ***Bandpass Filter*** **(BPF)**, which passes a contiguous band of frequencies not including DC and rejects the others. This is the most common type of filter in the payload. A BPF can be realized by a combination of a LPF and a HPF in cascade, but this is not common
- ***Band-Stop Filter***, which passes all frequencies but a contiguous band of frequencies that does not contain DC
- ***All-Pass Filter***, which passes all frequencies without attenuation.

The LPF is in a sense the most basic filter type. The reason is that the usual way to design a BPF is by "network synthesis," in which the first step is to define a LPF that meets the gain- and phase-response requirements, approximate it by a realizable network, and then transform the LPF to a BPF (Wikipedia, 2010a). A similar statement is true of a high-pass filter and a band-stop filter. The method of transformation is given in Wikipedia (2011a).

4.5.1.2 Filter-Response Families The filter responses that are by far most often used in the payload are the **Chebyshev**, **elliptic** or Cauer, and **pseudo- or quasi-elliptic**. The Chebyshev is probably the most used of these three because it is the easiest to implement (Holme, 2011).

The Laplace transfer functions of Chebyshev and elliptic filters (Appendix 4.A) are defined in terms of two different families of polynomials (Sorrentino and Bianchi, 2010). Both can have any positive-integer order. Both filters have "equiripple" gain responses, that is, they have a constant size of ripple, Chebyshev just in the passband and elliptic in both the passband and stopband, as shown in Figure 4.13. For the elliptic filter, the size of the two ripples can be different. There is also a related family of filter responses, the "generalized Chebyshev," for which the passband is equiripple. The two gain responses depicted in Figure 4.13 are fifth-order, which can be told by imagining the negative-ω side of the response, also. The peaks in the passband each come from a pole (see Appendix 4.A for a discussion on poles and zeros). The Chebyshev filter has only poles, while the elliptic filter has as many zeros as poles. The zeros cause the non-dB gain response to equal zero at a few frequencies out-of-band. They also cause the gain response to drop off more sharply outside the passband edge than the Chebyshev's does. For more information on filter responses for other orders, see, for example, Chang (1989).

The pseudoelliptic filter starts off as an elliptic design but the filter designer modifies it slightly to obtain the desired response. The polynomials belong to no

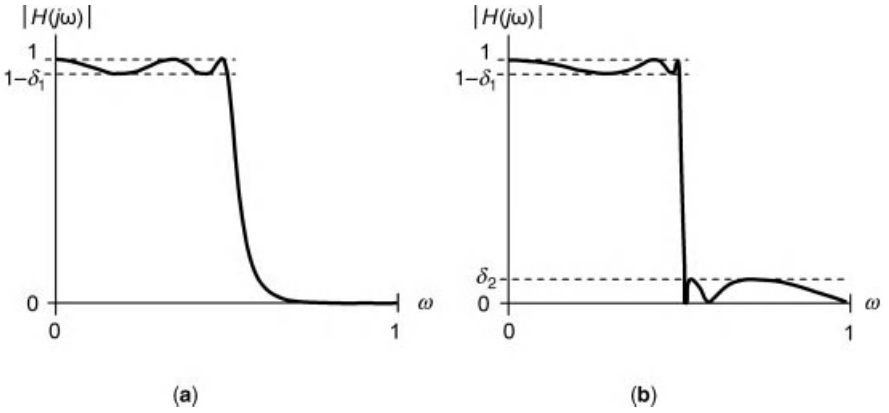


FIGURE 4.13 Gain response, not in dB, of 5th-order LPFs: (a) Chebyshev and (b) elliptic (after Wikipedia (2011b)).

particular class (Thompson and Levinson, 1988). A pseudoelliptic filter does not necessarily have the equiripple property (Kallianteris et al., 1977).

4.5.1.3 Resonator in General

A resonator in a payload filter has the shape of either a rectangular "solid" or a cylinder. The resonator boundary contains all or almost all of any electromagnetic fields in it. Theoretically, the resonator has an infinite number of resonant electromagnetic modes, and each mode corresponds to a particular resonant frequency. They are determined by the shape, dimensions, and dielectric constant of the material (or vacuum) of the resonator. If a signal is introduced into the resonator, the portion of the signal at these frequencies will excite the resonator, causing a standing wave. The mode with the lowest frequency is the **dominant** or fundamental mode. Because a resonator always has some RF losses, each resonant frequency is actually a narrow band of frequencies (Chang, 1989).

The naming convention for resonator modes is derived from the naming convention for modes of RF lines of the same general shape, rectangular or circular. A third index is added to the two indices of the RF-line mode. A resonator meant to resonate in only the dominant mode is a **single-mode** resonator. A **dual-mode** resonator is meant to resonate in two physically orthogonal modes at once. In this latter case, the modes are "degenerate," that is, they both have the same natural frequency. However, they can be tuned independently of each other, so they can realize separate poles (Fiedziuszko and Fiedziuszko, 2001). Each mode of a resonator creates one pole in the Laplace transfer function of the filter it will be part of.

A mode of a resonator has a tuning device, usually a screw that protrudes into the plane of the resonation. Adjusting the screw varies the electromagnetically effective dimensions of the resonator (Thompson and Levinson, 1988). An alternative to a screw for some filters is dent-tuning (Hsing et al., 2000).

A figure of merit is the (unloaded) quality or Q factor of a resonator:

$$Q = \frac{\omega_0 W}{P_\mathrm{d}}$$

where ω_0 is the resonant angular frequency, W is the total average energy stored in the resonator in one resonation period $2\pi/\omega_0$, and P_d is the average power dissipated in the resonator in one period (Sorrentino and Bianchi, 2010).

4.5.1.4 Filter in General

A filter is constructed from at least one resonator, the number depending on the number of poles desired in the filter and the number of resonance modes in each resonator. The resonators are coupled to each other by a mechanism which differs with the type of filter, as we will see. A filter has an input port and an output port.

The simplest arrangement of the resonators is in a line (called the "in-line," "longitudinal," "coaxial," or "axial" configuration in the literature), as illustrated in Figure 4.14 for a waveguide filter. Sequential resonators are coupled by irises between the cavities. This particular example of the in-line configuration has dual-mode resonators. That is why the irises have a cross-shape, so that two orthogonal

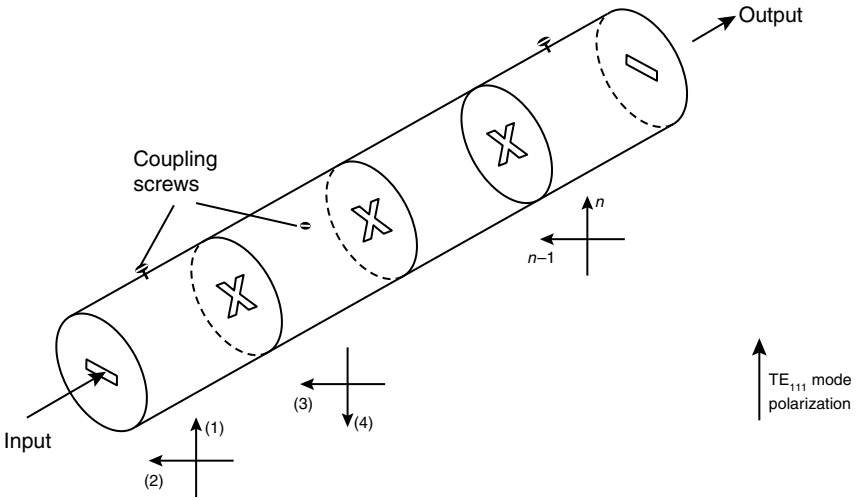


FIGURE 4.14 Circular-waveguide filter in in-line configuration, with mode-coupling screws. (©1972 IEEE. Reprinted, with permission, from Atia and Williams (1972)).

modes in one cavity can each couple to their corresponding modes in the adjacent cavity. An iris coupling one mode would be just a slot.

Coupling between nonsequential resonators, "cross-coupling," is often desirable, in order to modify the filter response. There are two types of modifications, and they can be made alone or simultaneously. One creates transmission zeros, usually for sharpening up the out-of-band gain rolloff. The other can equalize the phase response, potentially making the filter linear-phase, **self-equalized**. It is to be noted that a self-equalized filter does not have linear phase over the entire passband, just most of it. Which one of these two modifications a cross-coupling makes depends on the sign of the cross-coupling (Pfitzenmaier, 1982). In Figure 4.14, three mode-coupling screws are shown, where each screw couples the two orthogonal modes within one cavity. These are cross-couplings. The screws are at 45° from the plane of the input radiation, and the two different cases of 45° correspond to the two signs of cross-coupling.

A configuration that makes cross-coupling different cavities more mechanically feasible, is to place the resonators in two adjacent lines, with the resonator sequence taken to be along a U-shape from the input port, around the bend, and to the output port on the same end as the input port. An example of this "folded" or "planar" configuration is shown in Figure 4.15. This example has dual-mode cavities. Dual-mode irises are seen between cavities across from each other. There are other means for cross-coupling cavities diagonal from each other (not shown) (Atia and Williams, 1974).

An alternative way to create transmission zeros is the "extracted pole" technique (Holme, 2011). The "pole" here is an attenuation pole, that is, a transmission zero.

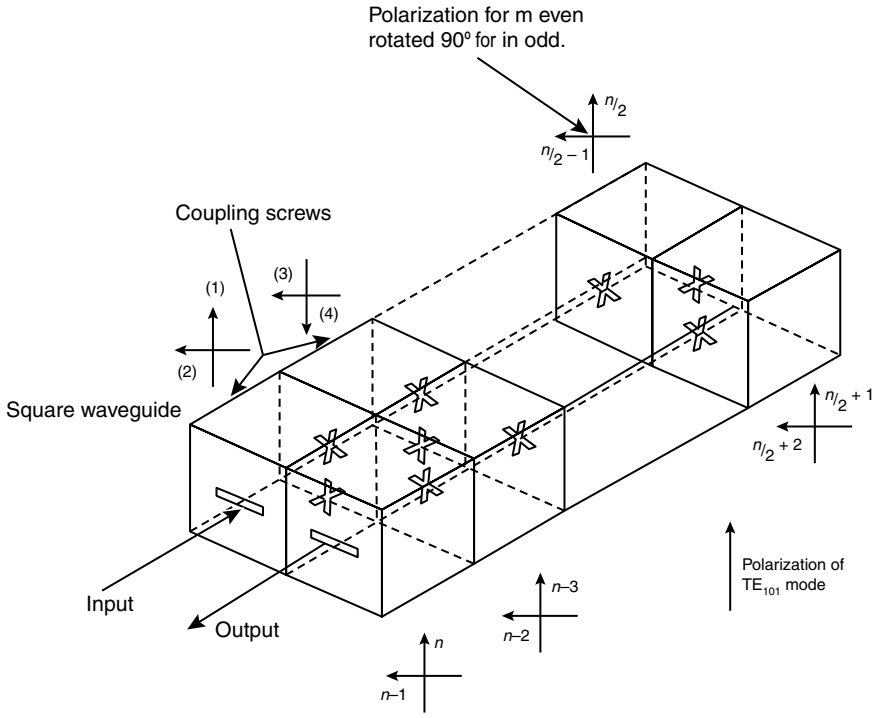


FIGURE 4.15 Rectangular-waveguide filter in folded configuration, with mode-coupling screws. (©1972 IEEE. Reprinted, with permission, from Atia and Williams (1972)).

The realization of the finite real poles (those on the $j\omega$-axis of the Laplace transfer function—see Appendix 4.A) is accomplished by extra cavities added before the input and after the output of the usual resonators in a folded configuration (Rhodes and Cameron, 1980).

The **Q factor of a filter** is that of its resonators. For a given filter technology, Q is inversely proportional to the filter's insertion loss (not in dB) (Kudsia et al., 1992). A Q factor of at least 8000 is usual for payload filters (Yu et al., 2003).

4.5.2 Filter Technology

4.5.2.1 Empty-Cavity Waveguide Filter We start our discussion of filters with waveguide filters. The other two filter technologies, resonator-cavity and combline, will be discussed in Sections 4.5.2.2 and 4.5.2.3, respectively. We discuss the resonator, the filter, and environmental aspects in the subsection for each filter technology.

The empty-cavity waveguide resonator is just what its name implies.

The filter made of such resonators may have either the in-line or the folded configuration as we have seen just above. The coupling mechanism between adjacent

TABLE 4.9 Properties of Waveguide Used in MUXes

	Mass Ranking (1st Best)	Effective CTE Ranking	Thermal Conductivity Ranking	Application
Aluminum, temperature-compensated	2nd	2nd (tied)	1st (tied)	IMUX and OMUX manifold OMUX BPFs
Invar, thin-wall, silver-plated	3rd (tied)	2nd (tied)	3rd (tied)	OMUX manifold IMUX and OMUX BPFs
Super Invar, silver-plated	3rd (tied)	1st	3rd (tied)	IMUX and OMUX BPFs
GFEC, silver-plated	1st	2nd (tied)	1st (tied)	OMUX BPFs

cavities is the iris. The coupling mechanism between orthogonal modes in one cavity is the screw.

The filters must meet their performance specifications over temperature. The materials used are various, and the choice is a trade-off among several considerations. The materials and their properties are summarized in Table 4.9 and stated in more detail here:

Aluminum, mechanically temperature-compensated. It is light weight and has high thermal conductivity, good for getting the heat out of the OMUX (Lundquist et al., 2002). It is about tied with Invar and GFEC (see below) in regard to the effective CTE of the BPFs made from it (Kudsia and O'Donovan, 1974). (Without compensation, the CTE of aluminum structures is about 15 times that of Invar (Wolk et al., 2002)). It is used for IMUX and OMUX manifolds and for OMUX BPFs (Tesat, 2007; Thales, 2006; Lundquist et al., 2002).

Invar, thin-wall, silver-plated. An alloy of mostly iron and 36% nickel (High Temp, 2011). Heavier than aluminum and GFEC. Its CTE is on a par with temperature-compensated aluminum and GFEC. It has low thermal conductivity (Lundquist et al., 2002). Used for OMUX manifolds and for IMUX and OMUX BPFs (Thales, 2006; Tesat, 2007).

Super Invar, silver-plated. An alloy of mostly iron, 32% nickel, and 5% cobalt (High Temp, 2011). Same weight as Invar. Same thermal conductivity (Thieberger et al., 2003). Best CTE of all (High Temp, 2011). Used for IMUX and OMUX BPFs (Fiedziuszko, 2002).

Graphite-fiber epoxy composite or graphite-fiber reinforced plastic (GFRP), silver-plated. Lightest weight (Fiedziuszko, 2002). Thermal conductivity about equal to aluminum's. Graphite BPFs are the most difficult to manufacture, so the most expensive (Kudsia et al., 1992). Used for OMUX BPFs (Fiedziuszko, 2002).

All the materials except Super Invar expand linearly with temperature.

A rather new method for combining a manifold of aluminum with BPFs of Invar (or other material) does not even require mechanical temperature-compensation of any element. The expansion of the aluminum of the manifold is not itself a problem because the manifold is very wideband. A deformable transition element is used between aluminum and Invar that accommodates the different expansion rates (Wolk et al., 2002).

Finally, there is a property of empty-cavity waveguide filters that one should keep in mind while studying test results. The filter's center frequency will shift on orbit relative to what it is on the ground. The reason is that waveguide's cutoff frequency is inversely proportional to the square root of the dielectric constant (Ramo et al., 1984), and vacuum and air have slightly different dielectric constants. The dielectric constant of vacuum is exactly 1.0, and that of air is 1.00059 at one atmosphere of pressure and 20°C (GSU, 2011). This means a frequency shift of about 300 ppm, which, for example, at 12 GHz is about 4 MHz.

4.5.2.2 Dielectric-Resonator Filter

The **dielectric-resonator loaded-cavity filter**, or simply **dielectric-resonator filter,** has much in common with the empty-cavity waveguide filter. They are much smaller than the corresponding waveguide filter would be (Fiedziuszko, 1982).

The main difference is that the resonators are pucks, that is, short cylinders, of temperature-stable dielectric material such as ceramic. The dielectric constant is 24–40, compared to 1 of vacuum. Twenty-four or so is good for higher frequencies due to the higher Q factor it provides (Holme, 2011). Each dimension of the resonator is reduced compared to an empty-cavity resonator, namely, divided by the square root of the dielectric constant (Zaki and Atia, 1983). The electromagnetic fields exist only in and in the near vicinity of the dielectric material (Fiedziuszko and Holme, 2001). Each resonator is placed in a waveguide cavity, sized so that the resonator's resonance frequencies are below the waveguide cutoff (Fiedziuszko, 1982). The waveguide cavity surrounding the puck does not resonate. The resonator is held in the middle of the cavity, that is, it does not touch the cavity walls (Fiedziuszko, 1982; Yu et al., 2004). The reason for keeping the resonator away from the waveguide walls is to reduce losses in the walls (Zaki and Atia, 1983). Fields are evanescent outside the resonator (Holme, 2011). The waveguide is there to prevent radiation of the fields and resultant degradation of resonator Q (Fiedziuszko and Holme, 2001).

The dielectric-resonator filters come in three different configurations, as shown in Figure 4.16, (a) in-line or longitudinal, (b) (linear) planar (sometimes also called "in-line"), and (c) folded (Tang et al., 1987). The in-line configuration is often used for dual-mode resonators and the folded for single-mode (Holme, 2011). Cross-coupling is done differently according to the configuration: (a) cross-shaped dual iris; (b) T-shaped dual iris; and (c) shaped slots as in Cameron et al. (1997).

Dielectric-resonator filters are able to achieve a Q factor of 8000 and beyond (Yu et al., 2003).

The temperature properties of a dielectric-resonator filter are basically those of the dielectric resonators. The dielectric material is made with a CTE lower than that

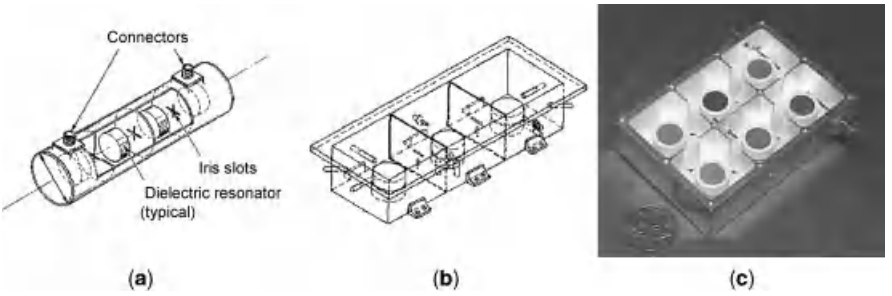


FIGURE 4.16 Dielectric-resonator filter configurations: (a) in-line, dual-mode (©1982 IEEE. Reprinted, with permission, from Fiedziuszko (1982)); (b) planar, dual-mode (©1992 IEEE. Reprinted, with permission, from Kudsia et al. (1992)); and (c) folded, single-mode, without lid (©2001 IEEE. Reprinted, with permission, from Fiedziuszko and Holme (2001)).

of Invar. Nothing special is required for the cavities (Fiedziuszko, 1982). The problem is to get the heat out of the dielectric resonators. One solution, for a filter with in-line arranged single- or dual-mode resonators, is to mount the resonator in its cavity on a support of thermally conductive, electrically insulating material (Holme et al., 1996). The trick here is to still accommodate inter-cavity irises.

4.5.2.3 Coaxial-Cavity Combline Filter The **coaxial-cavity combline filter** is used at the lowest frequencies, from L-band to C-band. It is different from the waveguide and dielectric-resonator filters, whose resonating modes are related to waveguide modes—this filter's resonators have TEM as dominant mode. This kind of filter is cheaper to build than the dielectric-resonator filter (Yu et al., 2003). It is easy to manufacture and tune (Yao and Atia, 2001).

There is confusing terminology around this filter. "Combline" is descriptive of a filter configuration, meaning that the longer-than-they-are-wide resonators are parallel and all attached on the same end to a grounding plate, looking like a comb. The combline configuration is closely related to the **interdigital** configuration, in which alternating resonators are attached to opposite grounding plates, looking like interlocking fingers. Not all combline filters look like a comb, though. Another complication is that there are other kinds of combline filter besides the coaxial cavity, one kind with resonators made of pieces of coax and another with dielectric rods in cavities. A "coaxial cavity" consists of a metal rod in a waveguide cavity, with vacuum as the dielectric between. Unlike in a dielectric-resonator filter, here the waveguide is an essential part of the resonator.

A coaxial cavity is much like stripline (Sorrentino and Bianchi, 2010). Both consist of a narrow conductor between two parallel, conducting ground-planes, as shown in Figure 4.17. For both, the electric field lines are perpendicular to the direction of propagation, as shown. The magnetic field lines, not shown, are perpendicular to the electric field lines and to the direction of propagation. The coaxial cavity is single-mode.

FIGURE 4.17 TEM-mode resonators and their electric field lines: (a) stripline (reprinted with permission of John Wiley & Sons, Ltd, from Sorrentino and Bianchi (2010)) and (b) coaxial cavity field lines (after Cristal (1964)).

A drawing of a coaxial cavity is given in Figure 4.18a. The center conductor is shorter than $\frac{1}{4}$th of the center wavelength, for example, $\frac{1}{6}$th (Yeung et al., 1993) and $\frac{1}{8}$th (Micro Comm, 2009). In fact, the center conductor is not a simple rod, because it must have a mechanism for frequency tuning. The capacitance of the unattached end of the rod, and thus the resonant frequency, is altered by means of a tuning screw inserted into the rod that has been hollowed out to some depth as illustrated in Figure 4.18b.

The coaxial-cavity combline filter consists of coupled TEM-mode transmission lines. Figure 4.19a shows the electric and magnetic fields of two adjacent resonators and how to couple the resonators. The total coupling has both magnetic and electric components, with the magnetic being stronger and the electric being out of phase with the magnetic. So the total coupling is the magnetic coupling minus the electric coupling. The magnetic coupling is decreased by a wall between the rods and the electric coupling is decreased by a screw. Two ways of cross-coupling resonators are shown in Figure 4.19b and c (Thomas, 2003).

The spacing between the parallel plates must be a small fraction of the filter's center wavelength, for example, 12%, for the filter truly to be a combline filter. As the spacing increases the resonator couplings are due not just to TEM but also partially to evanescent (i.e., below-cutoff) waveguide modes. This causes the Q factor to increase, for example, to double as the spacing goes from 12 to 20%, without

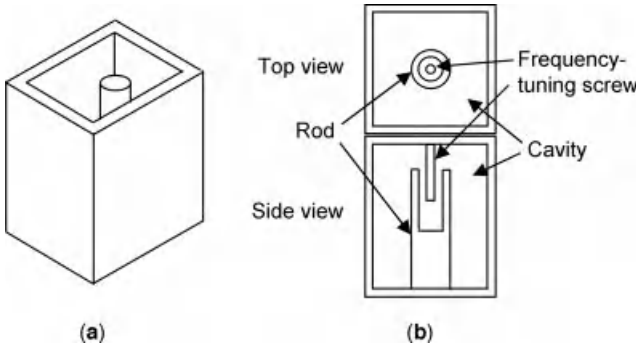


FIGURE 4.18 (a) Coaxial cavity and (b) showing frequency-tuning mechanism (after Yao and Atia (2001)).

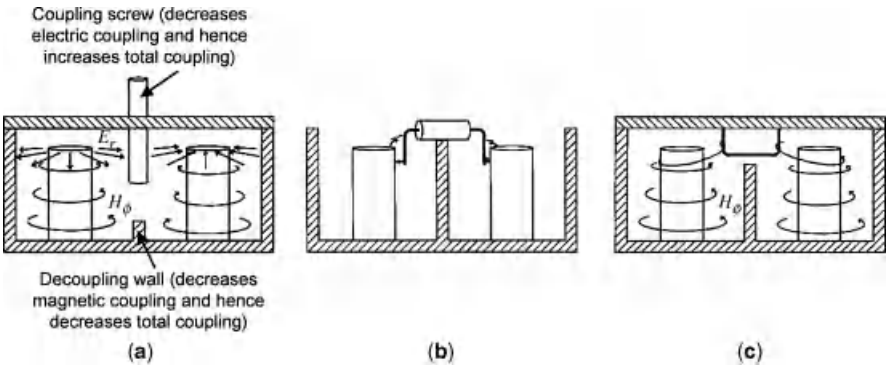


FIGURE 4.19 Coaxial-cavity resonators: (a) adjacent resonators, field lines, and coupling mechanisms; (b) capacitive coupling mechanism for nonadjacent resonators; and (c) inductive coupling mechanism for nonadjacent resonators. (©2003 IEEE. Reprinted, with permission, from Thomas (2003)).

increase in insertion loss. This does not happen to broadband combline filters or to the interdigital filter of any bandwidth (Levy et al, 1996).

Two combline filters of different configuration are shown without lids in Figure 4.20, (a) a noncoaxial-cavity combline filter, just to show that some combline filters look like a comb; and (b) a coaxial-cavity combline filter in a folded configuration, that looks like a brush. In both figures the frequency-tuning mechanisms are visible.

At C-band, a coaxial-cavity filter is smaller and lower in mass than a corresponding circular-waveguide filter would be (Smith and Yu, 2006).

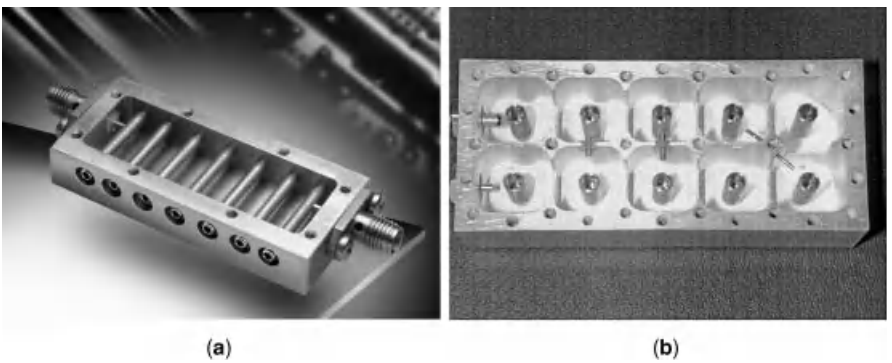


FIGURE 4.20 Combline filters without lids: (a) one that looks like a comb (not cavity type) (image courtesy of Link Microtek Ltd) and (b) one that looks like a brush (cavity type) (©2003 IEEE. Reprinted, with permission, from Yu et al. (2003)).

At L-band, a coaxial-cavity combline filter was able to achieve a Q factor of 2500, while a dielectric-resonator filter would not be (Yeung et al., 1993).

The coaxial-cavity filter by itself cannot achieve a Q of 8000. At C-band, the Q can be improved from 3,000 to well above 8,000 (20,000 was achieved) by moving the complex zeros of the Laplace-transform filter-response closer to the $j\omega$-axis by an iterative process. It does not alter the transmission zeros, that is, those on the $j\omega$-axis. This method is called "predistortion." It creates single-mode coaxial-cavity combline filters that are much smaller and lighter than single-mode dielectric-resonator filters. However, the insertion loss is much higher than that of the dielectric-resonator filter (Yu et al., 2003).

A low-power coaxial-cavity filter with cavities made of aluminum for low mass and rods made partly of Invar is very temperature-stable (Yao and Atia, 2001; Yu et al., 2003).

4.5.3 Filter Unit and Assemblies

The nonprocessing transponder that we are considering in this book is shown in Figure 4.21, with the filters outlined thickly. The preselect filter is a stand-alone unit, and we discuss it first. A transponder in general has multiplexers (MUXes), each of which either takes a signal and separates out the various channels from each other or combines various channels into one signal. The input multiplexer (IMUX) is the first MUX in a transponder and the output multiplexer (OMUX) follows it. We are assuming that the IMUX separates out channels and the OMUX combines channels. Since all MUXes share some characteristics, we discuss MUXes in general before going on to discuss separately the unique characteristics of IMUXes and OMUXes. We end by summarizing statements about technology performance.

4.5.3.1 Preselect Filter The preselect filter is the first filter unit that the signal sees, after the antenna and before the low-noise amplifier (LNA). If the antenna has a diplexer, the preselect filter may be part of the diplexer (Section 3.2.6). The preselect filter does two things to protect the LNA and the rest of the payload. One is to attenuate interfering signals that might saturate the wideband LNA and drive it into compression. Another thing is to attenuate interfering signals that may be stronger than the signals of interest, to ensure that the interferers are not passed by the IMUX channel filters.

The typical technology is coaxial-cavity combline, or interdigital, or waveguide (Kwok and Fiedziuszko, 1996). Coaxial-cavity combline are commercially available

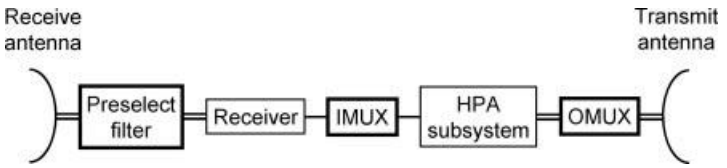


FIGURE 4.21 Simplified transponder diagram with filters outlined thickly.

for L-band to C-band (Micro Comm, 2009). The Q factor need not always be as high as 8000; for example, Q of 2500 with an insertion loss of 0.32 dB from coaxial-cavity combline technology was good enough for Inmarsat-3 at L-band (Yeung et al., 1993). Waveguide preselects are commercially available for X-band to Ka-band (CMT, 2011).

4.5.3.2 Multiplexer in General A MUX is actually an assembly of filters and payload-integration elements. The BPFs are "channel filters." They are most commonly dual-mode (Fiedziuszko, 2002). The MUX may contain other filters, too, for additional purposes.

Occasionally there is a situation where it makes sense to replace two channel filters over different bands by one filter with two passbands. This can be done by designing a filter whose band encompasses both passbands and then adding zeros to its response in the middle of the two passbands. The passband responses can be as desired. There can also be more than two passbands (Holme, 2002).

Of all the filters that a signal passes through in the payload (some of them are components of units), the only two that provide significant bandlimiting are the IMUX and OMUX. Together the IMUX and OMUX must provide channel selectivity. A question is how to divide the necessary amount of selectivity between the IMUX and the OMUX. The main reason that the IMUX and not the OMUX does most of the filtering is that the OMUX carries the post-HPA signal while the IMUX carries only a low-level signal. HPA power must not be wasted on unnecessary filter-insertion loss. The sharper a filter is, the higher its insertion loss. Also, narrow-band filters that can accommodate high power are difficult to make, so the fewer the OMUX resonators the better. Finally, means must be provided to carry away the heat in an OMUX filter resonator, so again the fewer the OMUX resonators the better. The consequence is that the IMUX BPFs are sharper, that is, have more poles, than the OMUX BPFs do, 8–10 poles versus typically 4–6 (Fiedziuszko, 2002; Garman et al., 2004).

4.5.3.3 Input Multiplexer The IMUX comes after the LNA. Two architectures are shown in Figure 4.22. The first one is the most common, the "channel-dropping

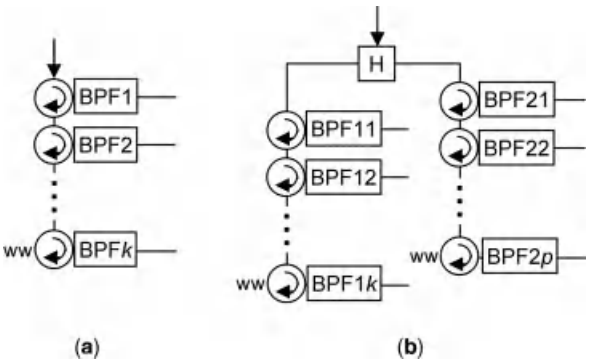


FIGURE 4.22 IMUX architectures: (a) channel-dropping and (b) channel-dropping with power splitter.

scheme" (Cameron and Yu, 2007). The signal from the LNA enters the first BPF. That filter passes one channel band and rejects the rest of the frequencies, which go back out the first circulator and through the second circulator to the second BPF. This continues until the frequencies rejected by all the BPFs get absorbed by the load on the last circulator. The farther down the chain it is, the larger the loss the channel band sees. The architecture shown in Figure 4.22b is similar to Figure 4.22a but has a power-divider that divides the signal into two copies, each copy going into a channel-dropping scheme (Thales, 2006). The loss distribution in this architecture is different from that of scheme (a). Both architectures depend on the LNA's gain being high enough to accommodate the IMUX loss.

The channels of an IMUX are rarely contiguous in frequency (Holme, 2011). The BPFs are self-equalized in regard to phase (Section 4.5.1.4) (Holme, 2011).

IMUXes are commercially available at least in waveguide and dielectric-resonator technologies: waveguide at K-band (Thales, 2006) and Ka-band (Garman et al., 2004) and dielectric-resonator at C-band to K-band (Thales, 2006).

4.5.3.4 Output Multiplexer The OMUX is the next filter assembly, usually after the HPAs. It combines channels amplified by different HPAs so that the combined signal can be transmitted on one polarization of an antenna beam. The OMUX is an assembly of BPFs (channel filters), a **manifold**, and other filters. A generic architecture is illustrated in Figure 4.23. A manifold is most often a length of waveguide short-circuited at one end (Fiedziuszko et al., 2002), to which the BPFs are attached at spacings of about one guide wavelength (Rhodes and Levy, 1979). A waveguide manifold is nearly lossless. The BPFs may be attached all on one side of the manifold ("comb" or "combline" arrangement), as shown, or on both sides ("herringbone" arrangement). At L-band, instead of a waveguide manifold, which would be very big, a coaxial 6-port junction has been used as a manifold. The BPFs are connected to it with semirigid coax (Fiedziuszko et al., 1989). Another manifold has been coaxial, and the BPFs are coupled to it (Lundquist et al., 2000).

Every channel filter should have good phase response. There are two ways to accomplish this. The much more common way is for the BPFs themselves to be self-equalized. The uncommon way is to add to the OMUX an external phase (or group-delay) equalizer and a circulator. The problem with this scheme, though, is

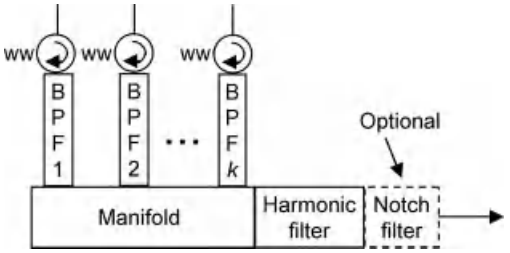


FIGURE 4.23 Typical OMUX architecture.

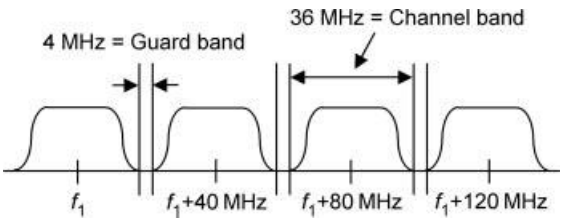


FIGURE 4.24 Example of contiguous channels separated by 10% guard bands.

that the circulator would be so sensitive to temperature that usually phase instability would arise (Holme, 2011).

The early OMUXes could only combine channels that were noncontiguous in frequency, that is, successive channels had to be separated by an empty band as wide as a channel. The first published report of a contiguous OMUX was by Chen et al. (1976). Later, more practical designs were presented. First came C-band in 1978. Then the more difficult Ku-band came, a five-channel one in 1982 and a 12-channel one, also by Chen, in 1983. (This history was provided by Holme (2011).) Contiguous channels separated by 10% guard bands are illustrated in Figure 4.24.

At the manifold output there is a **harmonic filter**, which attenuates the second and third harmonics of the HPA output as well as any HPA output in the payload's receive band (Saad, 1984). It is a low-pass filter. Also at the manifold output may be a **notch filter**, a very narrow-band-stop filter, to attenuate the near-out-of-band transmission in a protected frequency band. The OMUX architecture may be distributed, if it makes more sense for the particular payload architecture or spacecraft layout.

OMUXes are commercially available at least in waveguide and dielectric-resonator technologies: waveguide for X-band to K-band (Thales, 2006) and dielectric-resonator for L-band (Fiedziuszko et al., 1989; Lundquist et al., 2000) to C-band (Thales, 2006). For waveguide, 65 W per channel at C-band (Bosch, 1999) and 350 W at Ku-band (Seehorn et al., 2002) have been reported. For the dielectric-resonator BPFs, 50 W at L-band (Fiedziuszko et al., 1989) and 180 W at C-band (Thales, 2006) have been reported.

4.5.3.5 Summary of Filter-Technology Application

Table 4.10 summarizes the above statements of what filter technology is used for what commercial application.

4.5.4 Filter Specification

Table 4.11 gives an example of the parameters of a BPF specification. A parameter not seen for active units is the frequency shift over temperature in units of ppm per degree of temperature rise.

TABLE 4.10 Application of Commercial Filter Technologies

	Preselect	IMUX	OMUX	
	Frequency Bands	Frequency Bands	Frequency Bands	RF Power per Channel
Waveguide filter	X-band to Ka-band	K-band Ka-band	X-band to K-band	65 W at C-band 350 W at Ku-band
Dielectric-resonator filter	Unknown	C-band to K-band	L-band C-band	50 W at L-band 180 W at C-band
Coaxial-cavity combline filter	L-band to C-band	Unknown	Unknown	Unknown

TABLE 4.11 Example of Parameters in a BPF Specification

Parameter in BPF Specification	Units
Frequency range	GHz
Insertion loss	dB
Return loss—input, output	dB
Gain variation over band	dB pk-pk
Phase ripple over band	deg pk-pk
Power handling per channel (for OMUX only)	W
Frequency shift over temperature	ppm/°C
Operational temperature range	°C

4.6 SWITCH AND REDUNDANCY

4.6.1 Switch

Switches play a crucial role in the payload, since they allow changes in signal routing. They also allow selection of which units will be active out of a bank of units.

Various kinds of switch are used in payload integration, as shown in Figure 4.25. Some kinds are only made in coax, some only in waveguide, and two kinds in both. For the C-switch or transfer switch, the R-switch, and the T-switch, in some of their positions two signals flow through the switch. If two signals flow, the switch must be able to accommodate the RF power in the two signals combined. The R-switch has two positions in which two signals can flow through it, while all four of the T-switch positions have this feature. Two of the T-switch positions are equivalent. The T-switch paths are not all physically in one plane. Almost all switches are **latching**, meaning that the switch maintains a chosen RF contact path whether voltage is maintained or not after switching is accomplished.

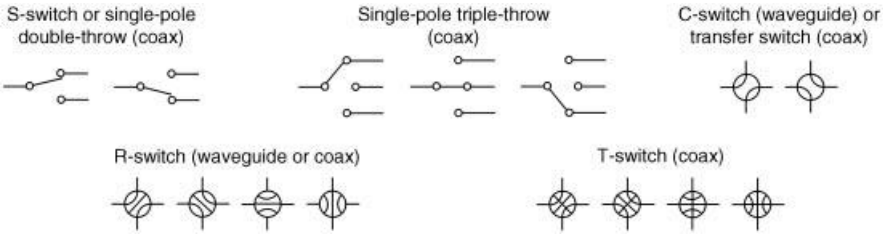


FIGURE 4.25 Switch types with all possible positions.

For coaxial switches, there are low- and high-power switches, where for one vendor low power means 5–10 W RF and high power means 33–102 W RF (Radiall, 2011a). Like coax itself, a coaxial switch works from DC to a GHz frequency. Its power-carrying ability decreases with increasing frequency. Insertion loss and VSWR increase with increasing frequency. Another vendor has a T-switch that can handle 260 W RF at 4 GHz (Dow-Key, 2011). This same vendor also makes switch modules, integrating several switches and perhaps power dividers into a block in order to provide redundancy switching at the payload input (Section 4.6.2).

Waveguide switches can in general accommodate higher RF power than coaxial switches can. One vendor makes modules of several ferrite-based S-switches (Com Dev, 2010). The switches have no moving parts; they are entirely solid-state for higher reliability. They are especially good for high-rate switching.

For especially crucial functions, switches themselves can be put in a redundant configuration. Figure 4.26 shows two examples in (a) a fully redundant one-input, two-output C-switch assembly and in (b) a partially redundant one-input, three-output R-switch assembly (Briskman and Prevaux, 2001). See next section for how the latter is used.

4.6.2 Redundancy

A bank of active units almost always includes spare, inactive unit(s), which can replace failed active unit(s). The units initially active on orbit are the **primary** units and the others the **spare or redundant** units. A bank of units contains either identical units, in the case of low-power units, or identical units differently tuned, in the

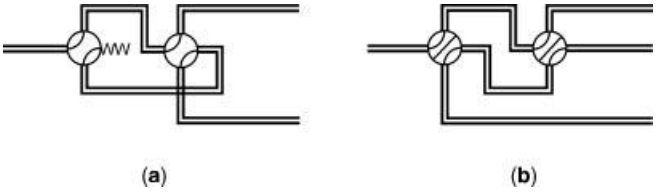


FIGURE 4.26 Redundant switches: (a) fully redundant, two-output C-switch assembly and (b) partially redundant, three-output R-switch assembly.

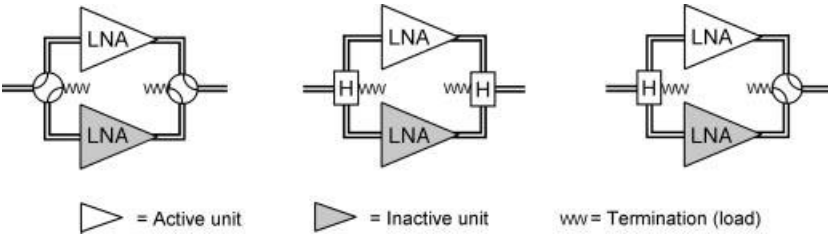


FIGURE 4.27 Redundancy via switches, hybrids, and mixture.

case of TWTA subsystems, where a TWTA subsystem is the TWTA plus its preamplifier.

Redundancy is described differently in the United States and Europe. For example, suppose that there are six units of which at any time four are to be active. In the United States, this is "6-for-4 redundancy" or simply "6 for 4" and written "6:4," while in Europe it may be written "4:2" or "4/6." The US terminology is used in this book.

Figure 4.27 shows that there is more than one way to access a spare unit when the primary unit fails. The most common way is with switches, in which case the switch collection and the unit bank together form a **redundancy ring**. However, hybrid couplers can also be used, or a mixture of hybrids and switches. The switches or hybrids need not be external to the units: in some cases of low-power units, a redundant unit and the switch or hybrid is a package with the primary unit.

We will see below that the RF line-and-switch structure in front of the unit bank is most often the reverse of the RF line-and-switch structure at the output of the unit bank. Additionally, the collection of switch positions at any time in front of the unit bank and the collection at the output have this property. Typically the switches are **ganged**, that is, when a redundancy-ring switch in front of the unit bank is commanded to turn, the corresponding switch on output automatically turns in the reverse direction.

Redundancy rings for the low-power units, that is, the LNAs and the frequency converters, usually have a different structure from the rings for the TWTAs (SSPA redundancy is discussed in Section 6.5.1.2). The low-power units are typically identical, or there may be a small number of banks in each of which the units are identical. The TWTAs, however, while perhaps identical in design, are typically optimized for different, narrow frequency bands. In addition, the low-power unit banks are typically small, while the TWTA banks can contain more than a hundred TWTAs.

Figure 4.28 gives four examples of redundancy rings for low-power units. The two examples in Figure 4.28a show how C- and T-switches can be used. The two in Figure 4.28b use the redundant switches introduced earlier.

For a TWTA ring, there are additional concepts and terms. Redundancy can be complete or partial. Suppose the redundancy is m:n. In **complete redundancy**, the failure of any m–n units can be accommodated by other units in the bank. In **partial redundancy**, there is some failure of m–n units that cannot be accommodated. The percentages of double failures, triple failures, and so on, up to and including m–n

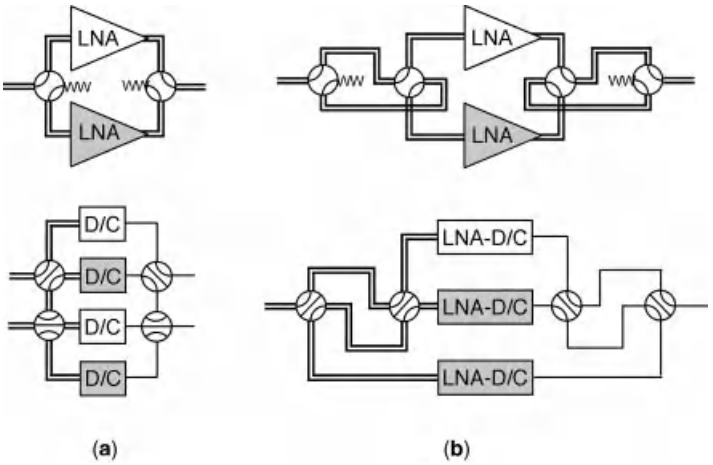


FIGURE 4.28 Low-power unit redundancy rings.

failures that can be accommodated, are all of interest in the calculation of the probability that the payload will meet its performance specification.

The **first-redundant** TWTA for a primary TWTA is the one that would be chosen in case the primary fails, assuming that no other has failed. The second-redundant is the unit that would be chosen if both primary and first redundant fail, assuming that no other has failed. The first-redundant TWTA for a band may be optimized for the same band as the primary, but the second redundant is typically optimized for a larger band that includes the band of the primary.

A redundancy ring for a TWTA bank of more than a few TWTAs has a regular or nearly regular structure, namely it is built up of blocks that are identical or similar. In each block are the primary and first-redundant units for each input line to the block, while the second-redundant units may be in a different block. The TWTAs in a block are connected, and the blocks are interconnected. How to construct rings with redundancy as close to complete as possible, with as few TWTAs as possible, with as few switches in paths (primary or redundant) as possible, with minimum interruption to existing signal paths when redundant units are switched in, and with provable redundancy has long been a subject of research. Recently an advance was made (Liang and Murdock, 2008). Figure 4.29 shows modular redundancy rings with the property that the first redundant can be switched in for any one TWTA and there is no interruption to the other paths. Figure 4.29a shows a module with 4:3 redundancy. All modules in this scheme are m:m–1. This module has three connections (vertical) to connect to other modules with. Figure 4.29b is an abstract representation of the module. Figure 4.29c shows how two such modules can be connected to form a small ring, shown abstractly in Figure 4.29d.

Figure 4.30 shows rings made of such four such modules, each module with 6:5 redundancy and each ring 24:20. Figure 4.30a and b shows two different fixed ways

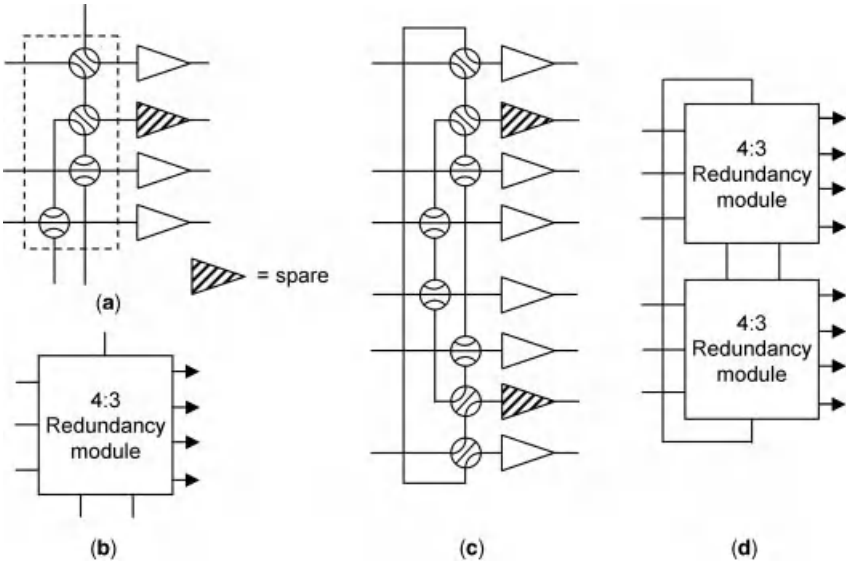


FIGURE 4.29 TWTA redundancy-ring modules, with noninterrupting first redundant (after Liang and Murdock (2008)).

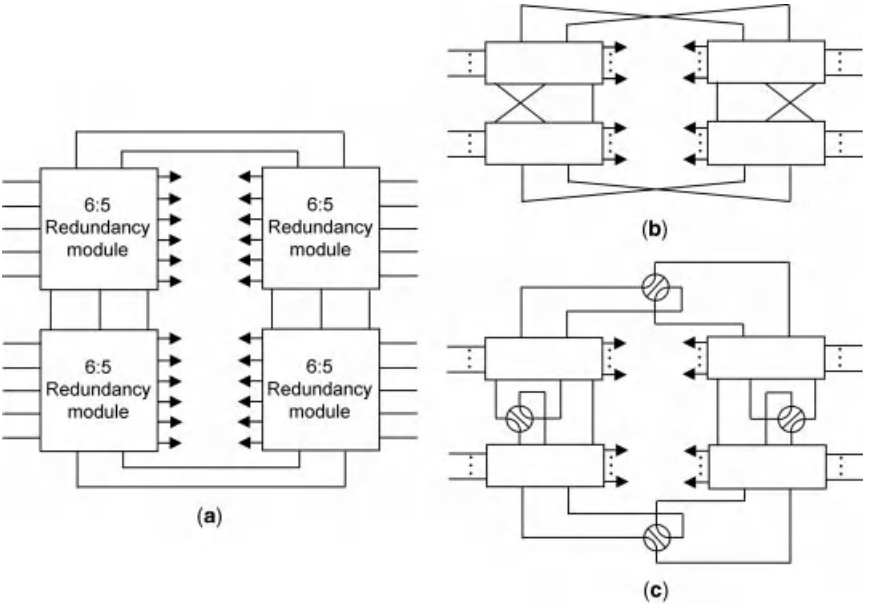


FIGURE 4.30 Integrated TWTA redundancy ring with noninterruptible modules, three different ways of connecting modules (after Liang and Murdock (2008)).

of interconnecting the modules. Each way allows partial 24:20 redundancy, covering different sets of failures. Figure 4.30c shows a switched interconnection method that is superior. The scheme is thought to provide complete redundancy when the number of modules is four or less. For any m:n, there is a way to constitute a ring of m–n modules that each includes one spare (Liang and Murdock, 2008).

APPENDIX 4.A

4.A.1 Filter Poles and Zeros

The usual way to design a BPF is by "network synthesis," in which the first step is to define a LPF that meets the requirements, approximate it by a realizable network thereby creating the "prototype filter," and then transform the prototype filter to passband (Wikipedia, 2010a). (A similar statement is true of high-pass and notch.) The method of transformation is given in Wikipedia (2011a). One exception to the kind of filter that can be designed this way is the all-pass filter for phase equalization.

A realizable LPF has a Laplace transfer function equal to the ratio of polynomials with real coefficients (Wikipedia, 2010b). Thus, the transfer function $H(s)$ of a realizable LPF has the following form:

$$H(s) = \frac{N(s)}{D(s)} \quad \text{where } N(s) \text{and } D(s) \text{are polynomials with real coefficients and}$$
$$s = \sigma + j\omega$$

A realizable filter has all its poles and zeros in the left half of the s-plane or on its boundary, the imaginary axis (Wikipedia, 2010b).

Suppose the degree of the numerator polynomial $N(s)$ is n and the degree of the denominator polynomial $D(s)$ is m. A polynomial has a number of roots equal to its degree. In general, not all the roots of a polynomial will be real. However, ones that are complex come in complex–conjugate pairs. The transfer function $H(s)$ can be written as follows:

$$H(s) = \frac{N(s)}{D(s)} = K\frac{(s-z_1)\ldots(s-z_n)}{(s-p_1)\ldots(s-p_m)} \quad \text{where } K \text{ is a real constant}$$

The roots z_i of $N(s)$ are the zeros of $H(s)$, and the roots p_k of $D(s)$ are the poles of $H(s)$. The order of a filter is the larger of n and m. A LPF that rolls off with increasing frequency, which is what we want, has a higher-degree polynomial on the bottom, namely, $m > n$. So the order of the filter is the degree of the denominator polynomial. It is useful to know that in the filter literature, a transmission zero is the same as an attenuation pole.

Figure 4.31 shows the **pole-zero diagram** or plot for the fourth-order low-pass Chebyshev and elliptical filters. The Chebyshev filter has four poles (actually, it also

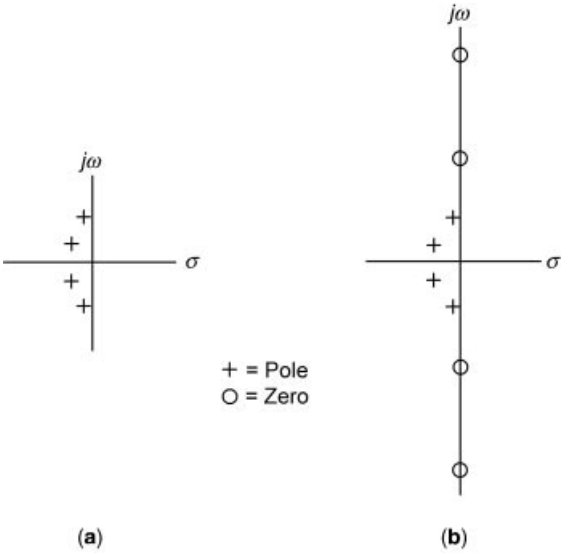


FIGURE 4.A.1 Pole-zero diagram for 4th-order (a) Chebyshev filter and (b) elliptic filter (after Maxim (2008)).

has four zeros but they are at infinity). The elliptic filter has an equal number of poles and zeros (all finite), which is a property of elliptic filters (Maxim, 2008). The elliptic filter poles are not quite in the same place as the Chebyshev filter poles.

REFERENCES

Aeroflex Weinschel (2010). Fixed coaxial attenuators. Catalog. Jan. On www.aeroflex.com/AMS/Weinschel/PDFiles/fixedatten.pdf. Accessed Sept. 30, 2011.

Atia AE and Williams AE (1972). Narrow-bandpass waveguide filters. *IEEE Transactions on Microwave Theory and Techniques;* 20; 258–265.

Atia AE and Williams AE (1974). Nonminimum-phase optimum-amplitude bandpass waveguide filters. *IEEE Transactions on Microwave Theory and Techniques;* 22; 425–431.

Bosch Telecom GmbH (now Tesat-Spacecom GmbH) (1999). Passive components for space applications. Product brochure. June. On www.electronicnote.com/media/downloads/Bosch%20Passive%20Components.pdf. Accessed Nov. 19, 2011.

Briskman RD and Prevaux RJ (2001). S-DARS broadcast from inclined, elliptical orbits. International Astronautical Congress; Oct. 1–5; pp. 503–518.

Cameron RJ, Tang W-C, and Dokas V (1997). Folded single mode dielectric resonator filter with cross couplings between non-sequential adjacent resonators and cross diagonal couplings between non-sequential contiguous resonators. U.S. patent 5,608,363. Mar. 4.

Cameron RJ and Yu M (2007). Design of manifold-coupled multiplexers. *IEEE Microwave Magazine;* Oct.; 46–59.

Chang K, editor (1989). *Handbook of Microwave and Optical Components, Vol. 1, Microwave Passive and Antenna Components*, New York: John Wiley & Sons, Inc.

Chen MH, Assal F, and Mahle C (1976). A contiguous band multiplexer. *COMSAT Technical Review;* 6; fall; 285–307.

CMT Inc (2011). Waveguide products. Entry page to product information on waveguide filters. Sept. On www.cmtfilters.com/waveguide.php. Accessed Nov. 20, 2011.

Cobham PLC, Continental Microwave Division (2006). *Waveguide Component Specifications and Design Handbook*, 7th ed., Nov. On www.cobham.com/about-cobham/aerospace-and-security/about-us/antenna-systems/exeter/products-and-services/passive-rf-components.aspx. Accessed Sept. 15, 2011.

Com Dev Europe (2010). Ferrite based RF switches. Product brochure. Mar. On www.com-devintl.com/default.aspx?id=260. Accessed Sept. 28, 2011.

Cristal EG (1964). Coupled circular cylindrical rods between parallel ground planes. *IEEE Transactions on Microwave Theory and Techniques;* 12; (Jul.); 428–439.

Dow-Key Microwave (2011). Space products. Product brochure. On www.dowkey.com/_upload/0/SpaceSection_spread.pdf. Accessed Sept. 28, 2011.

Fiedziuszko SJ (1982). Dual-mode dielectric resonator loaded cavity filters. *IEEE Transactions on Microwave Theory and Techniques;* 30; 1311–1316.

Fiedziuszko SJ (2002). Satellites and microwaves. *International Conference on Microwaves, Radar and Wireless Communications;* 3 (May 20–22); 937–953.

Fiedziuszko SJ, Doust D, and Holme S (1989). Satellite L-band output multiplexer utilizing single and dual mode dielectric resonators. *IEEE MTT-S International Microwave Symposium Digest;* 2 (June 13–15); 683–686.

Fiedziuszko SJ and Holme S (2001). Dielectric resonators, raise your high-Q. *IEEE Microwave Magazine;* Sept.; 51–60.

Fiedziuszko SJ and Fiedziuszko GA, inventors; Space Systems/Loral, Inc, assignee (2001). General response dual-mode dielectric resonator loaded cavity filter. U.S. patent 6,297,715 B1. Oct. 2.

Fiedziuszko SJ, Holme SC, and O'Neal NL, inventors; Space Systems/Loral, Inc, assignee (2002). Microwave multiplexer with manifold spacing adjustment. U.S. patent 6,472,951 B1. Oct. 29.

Garman SM, Kaliski MAR, and Thorburn MA (2004). SS/L's Ka-band technology enables broadband access via satellite. *AIAA International Communications Satellite Systems Conference;* May 9–12; 1–4.

Georgia State University, Dept. of Physics and Astronomy (2011). Dielectric constants at 20°C. On hyperphysics.phy-astr.gsu.edu/hbase/tables/diel.html. Accessed Oct. 17, 2011.

High Temp Metals, Inc (2011). Product technical data. On www.hightempmetals.com/technicaldata.php. Accessed Oct. 20, 2011.

Holme S (2002). Multiple passband filters for satellite applications. *AIAA International Communications Satellite Systems Conference;* May 12–15; 1–4.

Holme SC, Space Systems/Loral (2011). Private communication. Nov. 8.

Holme SC, Fiedziuszko SJ, and Honmyo Y, inventors; Space Systems/Loral, Inc, assignee (1996). High power dielectric resonator filter. U.S. patent 5,515,016. May 7.

Hsing CL, Jordan JE, and Tatomir PJ, inventors; Hughes Electronics Corp, assignee (2000). Methods of tuning and temperature compensating a variable topography electromagnetic wave device. U.S. patent 6,057,748. May 2.

Kallianteris S, Kudsia CM, and Swamy MNS (1977). A new class of dual-mode microwave filters for space application. *European Microwave Conference;* Sept. 5–8; 51–58.

Kudsia C, Cameron R, and Tang W-C (1992). Innovations in microwave filters and multiplexing networks for communications satellite systems. *IEEE Transactions on Microwave Theory and Techniques;* 40; 1133–1149.

Kudsia CM and O'Donovan MV (1974). A light weight graphite fiber epoxy composite (GFEC) waveguide multiplexer for satellite application. *European Microwave Conference;* Sept. 10–13; 585–589.

Kwok RS and Fiedziuszko SJ (1996). Advanced filter technology in communication satellite systems. *Proceedings of International Conference on Circuits and System Sciences;* June 20–25; 1–4. On www.engr.sjsu.edu/rkwok/Publications/Adv_Filter_Proc_ICCASS_p155_1996.pdf. Accessed Oct. 11, 2011.

Levy R, Yao H-W, and Zaki KA (1996). Transitional combline/evanescent mode microwave filters. *IEEE MTT-S International Microwave Symposium Digest;* 2 (June 17–21); 461–464.

Liang S and Murdock G (2008). Integrated redundancy ring based on modular approach. *AIAA International Communications Satellite Systems Conference;* June 10–12; 1–7.

Lundquist S, Mississian M, Yu M, and Smith D (2000). Application of high power output multiplexers for communications satellites. *AIAA International Communications Satellite Systems Conference;* 1 (Apr. 10–14); 1–9.

Lundquist S, Yu M, Smith D, and Fitzpatrick W (2002). Ku-band temperature compensated high power multiplexers. *AIAA International Communications Satellite Systems Conference;* May 12–15; 1–7.

Maxim Integrated Products, Inc (2008). A filter primer. Application note 733. Oct. 6. On www.maxim-ic.com/app-notes/index.mvp/id/733. Accessed Oct. 10, 2011.

Mele S, Cobham PLC, Exeter, New Hampshire (2011). Private communication. Sept. 19.

Micro-Coax, Inc (2001). Space capabilities. Brochure. Feb. On www.micro-coax.com/pages/products/pdfs/Spacecapabilities.pdf. Accessed Sept. 23, 2011.

Micro-Coax, Inc (2008). Microwave & RF Cable. Catalog. Sept. 4. On www.micro-coax.com/pages/products/pdfs/MCCBulkCableNC.pdf. Accessed Sept. 23, 2011.

Micro Communications, Inc (2009). Microwave products. Product literature. Mar. On www.mcibroadcast.com/file/RYMSA%20Brochurev3.pdf. Accessed Nov. 12, 2011.

National Institute of Standards and Technology, US, Mechanical Metrology Division (2004). Engineering metrology toolbox, temperature tutorial. Nov. 9. On emtoolbox.nist.gov/Temperature/Slide14.asp. Accessed Sept. 20, 2011.

Pfitzenmaier G (1982). Synthesis and realization of narrow-band canonical microwave bandpass filters exhibiting linear phase and transmission zeros. *IEEE Transactions on Microwave Theory and Techniques;* 30; 1300–1311.

Radiall S.A. (2011a). RF & Microwave Coaxial Switching Products 2011. On www.radiall.com/radiall/easysite/Radiall/products/catalogs-download. Accessed Sept. 27, 2011.

Radiall S.A. (2011b). Space qualified phase shifter. Product page. On www.radiall.com/radiall/easysite/Radiall/products/space/phase-shifter. Accessed Sept. 27, 2011.

Ramo S, Whinnery JR, and Van Duzer T (1984). *Fields and Waves in Communication Electronics*, 2nd ed., New York: John Wiley & Sons, Inc.

Rhodes JD and Cameron RJ (1980). General extracted pole synthesis technique with applications to low-loss TE_{011} mode filters. *IEEE Transactions on Microwave Theory and Techniques;* 28; (Sept.); 1018–1028.

Rhodes JD and Levy R (1979). Design of general manifold mutiplexers. *IEEE Transactions on Microwave Theory and Techniques;* 27; 111–123.

Saad AMK (1984). Novel lowpass harmonic filters for satellite application. *IEEE MTT-S International Microwave Symposium Digest;* May 30–June 1; 292–294.

Schennum GH, Space Systems/Loral (2011). Private communication. Oct. 24.

Seehorn R, Ziegler C, Holme S, and Fiedziuszko G (2002). High power high temperature Ku band multiplexers. *AIAA International Communications Satellite Systems Conference;* May 12–15; 1–4.

Smith DJ and Yu M (2006). Microwave resonator and filter assembly. U.S. patent 7,075,392 B2. Jul. 11.

Sorrentino R and Bianchi G (2010). *Microwave and RF Engineering.* Chichester, UK: John Wiley & Sons, Ltd.

Tang W-C, Siu D, Beggs BC, and Sferrazza J, inventors; Com Dev Ltd, assignee (1987). U.S. patent 4,652,843. Mar. 24.

Teledyne, Storm Products Co (2010). *Microwave: high performance interconnect products*. Apr. On www.stormproducts.com/resource.asp. Accessed Sept. 16, 2011.

Tesat-Spacecom GmbH (2007). Facts, products and services: passive RF products. Product-line brochure. On www.tesat.de/index.php?option=com_content&view=article&id=11&Itemid=16&lang=en. Accessed Dec. 23, 2008.

Thales Alenia Space (2006). IMUX & OMUX. Product datasheets. On www.thalesgroup.com/Pages/Solution.aspx?id=3435&pid=1500. Accessed Jan. 27, 2011.

Thieberger P, Kirk HG, Weggel RJ, and McDonald K (2003). Moving solid metallic targets for pion production in the Muon Collider/Neutrino Factory project. *Proceedings of Particle Accelerator Conference;* 3 (May 12–16); 1634–1636.

Thomas JB (2003). Cross-coupling in coaxial cavity filters—a tutorial overview. *IEEE Transactions on Microwave Theory and Techniques;* 51 (Apr.); 1368–1376.

Thompson JD and Levinson DS, inventors; Hughes Aircraft Co, assignee (1988). *Microwave directional filter with quasi-elliptic response*. U.S. patent 4,725,797. Feb. 16.

Wikipedia (2010a). Network synthesis filters. Aug. 19. On en.wikipedia.org. Accessed Oct. 14, 2011.

Wikipedia (2010b). Positive-real function. Nov. 27. On en.wikipedia.org. Accessed Oct. 14, 2011.

Wikipedia (2011a). Prototype filter. Jul. 13. On en.wikipedia.org. Accessed Oct. 10, 2011.

Wikipedia (2011b). Chebyshev filter. Sept. 14. On en.wikipedia.org. Accessed Sept. 14, 2011.

Wikipedia (2011c). PTFE. Sept. 15. On en.wikipedia.org. Accessed Sept. 25, 2011.

W. L. Gore & Associates (2003). Gore spaceflight microwave cable assemblies. Feb. On www.gore.com/MungoBlobs/883/162/gore_spaceflight_microwave_cable_assemblies_catalog.pdf. Accessed Sept. 23, 2011.

Wolk D, Damaschke J, and Schmitt D, inventors; Robert Bosch GmbH, assignee (2002). Frequency-stabilized waveguide arrangement. U.S. patent 6,433,656 B1. Aug. 13.

Yao H-W and Atia AE (2001). Temperature characteristics of combline resonators and filters. *IEEE MTT-S International Microwave Symposium Digest;* 3 (May 20–25); 1475–1478.

Yeung TK, Gregg H, and Morgan I (1993). Lightweight low noise space qualified L-band LNA/filter assemblies. *European Conference on Satellite Communications;* Nov. 2–4; 122–127.

Yu M, Smith D, and Ismail M (2004). Half-wave dielectric rod resonator filter. *IEEE MTT-S International Microwave Symposium Digest;* 2 (June 6–11); 619–622.

Yu M, Tang W-C, Malarky A, Dokas V, Cameron R, and Wang Y (2003). Predistortion technique for cross-coupled filters and its application to satellite communication systems. *IEEE Transactions on Microwave Theory and Techniques;* 51; (Dec.); 2505–2515.

Zaki KA and Atia AE (1983). Modes in dielectric-loaded waveguides and resonators. *IEEE Transactions on Microwave Theory and Techniques;* 31; 1039–1045.

CHAPTER 5

LOW-NOISE AMPLIFIER AND FREQUENCY CONVERTER

5.1 INTRODUCTION

This chapter is about the low-noise amplifier (LNA) and the frequency converter, two units in the front end of the payload. We discuss their architecture, redundancy scheme, and environmental considerations and give an example of their communications-related specification parameters. We also explain some of the parameters along the way.

The LNA sets, or to a large extent sets, the noise figure of the payload (Section 7.12.1). Its essential feature is that it has a low noise figure and high gain. Any resistive losses before the LNA add one-to-one to the payload noise figure in dB, so the LNA must be placed as close as possible to the antenna terminal.

The frequency converter changes the carrier frequency of the signal. Uplink and downlink frequencies are always different. Most commonly for nonprocessing payloads, there is only one frequency conversion, a downconversion (D/C). However, if the uplink and downlink frequencies are higher than Ku-band, there may be a downconversion to an IF at which the signal is filtered, followed by an upconversion (U/C). For regenerative payloads, the D/C would be to baseband, followed by an U/C. The D/C is almost always located right next to the LNA to reduce the losses on the following transmission line to the IMUXes, since lower frequencies have less loss in waveguide and coaxial cable.

The rest of this chapter is organized as follows:

- *Section 5.2*: LNAs and frequency converters at the level of payload architecture

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

- *Section 5.3*: Intermodulation products of interest for the LNA and the frequency converter
- *Section 5.4*: LNA: unit architecture, technology, nonlinearity, environmental considerations, and specification
- *Section 5.5*: Frequency converter: conversion architecture, introduction to phase noise, converter unit architecture, mixer, reference oscillator, local oscillator, nonlinearity, environmental, and specification
- *Appendix 5.A*: Formula for integrating phase-noise spectrum.

5.2 LOW-NOISE AMPLIFIERS AND FREQUENCY CONVERTERS IN PAYLOAD

5.2.1 Architecture in Payload

A LNA and a D/C together are sometimes called a **receiver**. This name makes sense when each active LNA is paired with one active D/C, which is a common architecture in nonprocessing payloads. However, other configurations are possible, as illustrated in Figure 5.1. Figure 5.1a shows a one-LNA-to-multiple-D/C architecture, (b) a one-to-one architecture, and (c) a multiple-LNA-to-one-D/C architecture. The Figure 5.1a architecture is also common among nonprocessing payloads. The choice between (a) and (b) depends primarily on how the uplink and downlink channels are related. The Figure 5.1c architecture is sometimes used with an active phased-array antenna, where in this case the antennas in the figure represent antenna-array elements and the D/C output a downconverted beam.

5.2.2 Redundancy Scheme

Figure 5.2 shows various redundancy schemes for two active receiver channels. Redundancy schemes for the LNA or the D/C separately can easily be deduced. All four schemes have four LNAs and four D/Cs. Scheme (a) provides the lowest reliability, schemes (b) and (c) provide higher reliability than (a), and scheme (d) provides the highest. This does not necessarily mean that (b) is better than (a), and so on. If the reliability provided by (a) is high enough when considered as a contributor

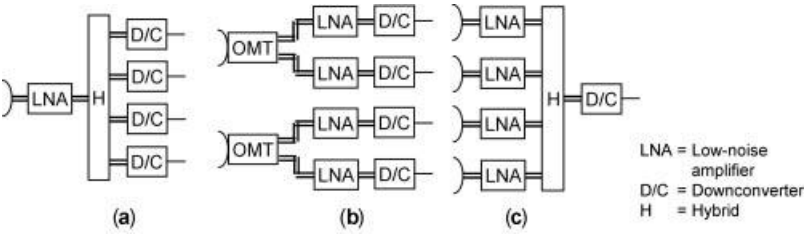


FIGURE 5.1 Examples of LNA-D/C architecture in payload (redundancy not shown).

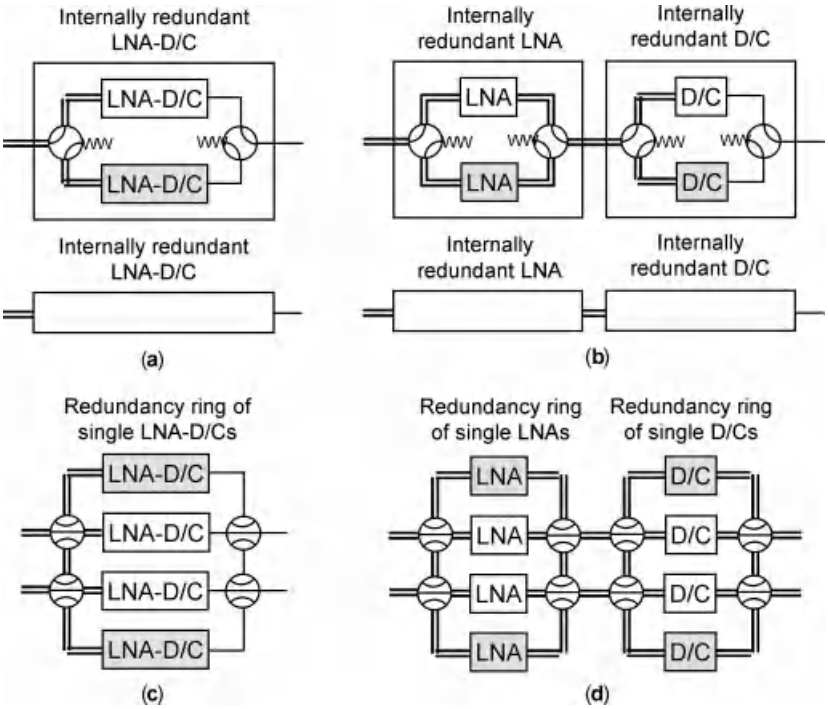


FIGURE 5.2 Receiver redundancy schemes for two active channels.

to the overall payload reliability then it is unnecessary and wasteful to employ the additional switches and waveguide of the other schemes.

Redundancy of 2:1 or $2k$:k for integer k is most common for the LNA, but the Sirius Radio satellite system has a switched 3:1 redundancy ring for its single active receiver chain (Briskman and Prevaux, 2001).

5.2.3 Combining

LNA or receiver outputs are not normally combined but they are in an active, receive phased array. Let us suppose we are dealing with LNAs, but the following would equally as well apply to receivers. In such an array, each array element is connected to its own LNA. To form each beam, all the LNA outputs must be properly phased and perhaps gain-controlled before being combined, so phase and gain stability over temperature and life of the LNAs and beam-forming network are crucial. The LNAs must individually exhibit stability, and the difference between any pair over a limited temperature range and life must exhibit stability (Yeung et al., 1993). The lower the frequency band is, the longer the guided wavelength, which makes it easier to achieve stability in the beam-forming network.

5.3 INTERMODULATION PRODUCTS

To understand what the amplifier and frequency converter do, we must take a brief detour to learn about the **intermodulation product (IMP)**.

Amplifiers and frequency converters are inherently nonlinear elements, meaning that their response to a signal cannot be represented by a filter. A filter output can be thought of as a linear combination of infinitesimally short time segments of the input signal or of infinitesimally narrow frequency segments of the input signal. Specifically, the LNA is nonlinear because the amplification and phase shift it provides depend on the power of the input RF signal; the frequency converter is nonlinear because it changes the frequency band.

When two tones combine in a nonlinear element, the output is many tones. If the tone frequencies are f_1 and f_2 in Hz then the **frequency** and **order** of the (m,n)th IMP are given by

$$\text{IMP}(m,n) \text{ has frequency} = |mf_1 - nf_2| \text{ and order} = |m| + |n| \text{ for integers } m \text{ and } n$$

Note that IMP(m,n) and IMP($-m$,$-n$) are identical. The outputs at f_1 and f_2 are at the **fundamental frequencies**; the outputs at multiples of f_1 and multiples of f_2 are **harmonics**; and the other outputs are **cross-products**. Figure 5.3 gives an example of the IMPs through 4th order when f_1 and f_2 are close and $f_1 < f_2$. The DC IMPs are not shown because RF devices do not couple to DC. The IMPs' magnitudes are not shown, either, since they depend on the magnitudes of the fundamentals and the characteristics of the active element. Figure 5.4 shows another example of the same IMPs but for two tones not spaced close together. Figure 5.4a is primarily an example of two tones being amplified together, while Figure 5.4b is primarily an example of a RF or IF tone and a local oscillator (LO) mixing in a

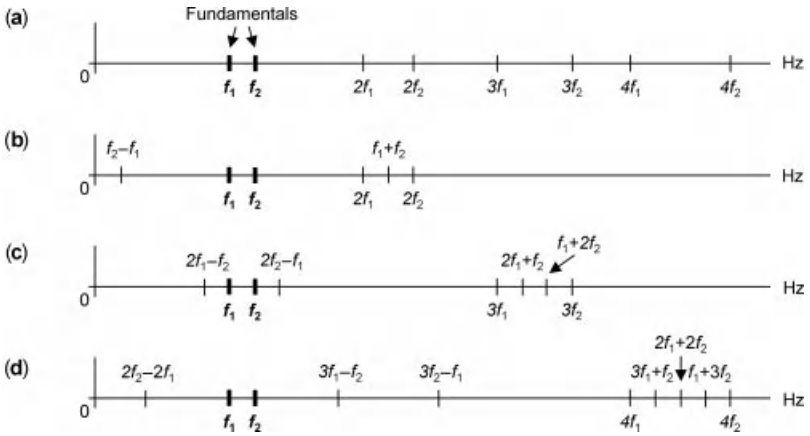


FIGURE 5.3 IMPs of two tones close together: (a) harmonics through 4th-order, (b) 2nd-order products, (c) 3rd-order products, and (d) 4th-order products.

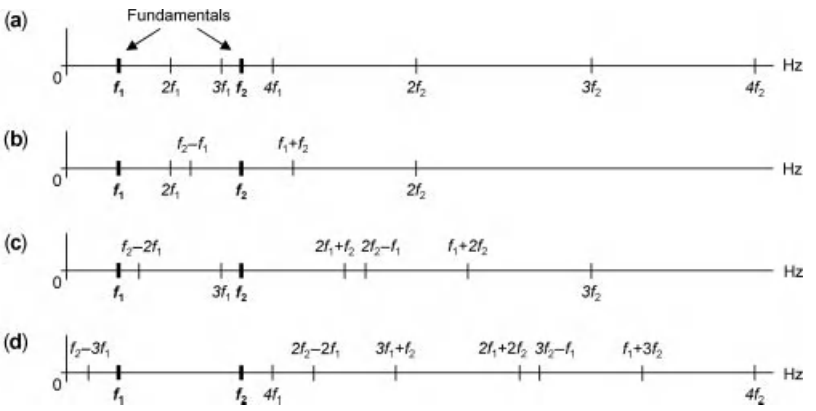


FIGURE 5.4 IMPs of two tones not close together: (a) harmonics through 4th-order, (b) 2nd-order products, (c) 3rd-order products, and (d) 4th-order products.

frequency downconverter. The point of having these two examples is to show that depending on the frequency spacing different IMPs may be a problem, for example, be inband or interfere significantly into another channel.

Other numbers of signals besides two can be input to a nonlinear element. A one-tone input yields output at the fundamental and in general all harmonics. A three-tone input at f_1, f_2, and f_3 yields IMPs in general at all frequencies of the form $|mf_1 - nf_2 + kf_3|$, and similarly for any number of tones input. A modulated-signal input can be thought of as consisting of many tones very close together; then the output of the nonlinear element can be roughly conceived of. There is more discussion on the IMPs of modulated signals in Section 8.4.2.1.

For an amplifier, the IMPs of interest derive from tones inband or near-inband. For a frequency converter, two sets of IMPS are of interest: (1) those deriving from the mixing of an RF or IF input tone with the LO (see Section 5.5.1 below) and (2) those deriving from two or more RF or IF input tones.

5.4 LOW-NOISE AMPLIFIER

5.4.1 LNA Unit Architecture and Technology

The LNA consists of a single chain of component amplifiers. The first-stage amplifier has the lowest noise figure, while the last-stage amplifier has the greatest gain.

Because the LNA's noise figure is the basis of the entire payload's noise figure, a manufacturer with the lowest noise figure has a key advantage, work is always going on to improve low-noise transistor technology. A person involved with the payload may desire at least a beginning-level familiarity with the technology. For basic information on transistor and IC technology, two encyclopedic web sites are recommended (Wikipedia and Microwave Encyclopedia).

There are currently three state-of-the-art technologies for the low-noise transistor in payload LNAs, GaAs pHEMT, InP HEMT, and GaN HEMT. GaAs is a gallium–arsenide compound semiconductor, InP an indium–phosphide compound semiconductor, and GaN a gallium–nitride compound semiconductor. A pHEMT is a pseudomorphic HEMT, and a HEMT is a high-electron-mobility transistor. All of these transistors are built on or grown on top of a substrate. They have a gate, whose short dimension when you look at it, the gate length [*sic*], has a major effect on the maximum operational frequency (Microwave Encyclopedia, 2006). The gate length is sometimes stated immediately before the transistor technology, for example, in "0.15 μm GaAs pHEMT." Sometimes the compound semiconductors used in the layers above the substrate are stated, too, for example, in "AlInAs/GaInAs HEMT on InP substrate" or "InP-based InGaAs/InAlAs HEMT." The GaN HEMT is sometimes called "GaN FET" or "GaN HFET," where a FET is a field-effect transistor and a HFET is a heterojunction FET.

The oldest of the three technologies, the GaAs pHEMT, is used for GHz frequencies up through about 30 GHz. This pHEMT has a GaAs substrate. The device is made in discrete form as well as in monolithic microwave integrated-circuit (MMIC) form (see, e.g., Sowers et al. (2001)). In recent years, the InP HEMT has at least partly taken over at 30 GHz and is the lowest-noise technology for frequencies up through 100 GHz and even higher (Chou et al., 2003). This HEMT has an InP substrate. At first only discrete devices were made in the technology but since 2000 MMICs have been made (Paine et al., 2000). The newest technology is the GaN HEMT, currently used for low-noise applications up to about 16 GHz (Micovic et al., 2007). Contrary to the naming convention of the other devices, the GaN HEMT has a SiC (silicon carbide) substrate and a GaN channel. It has a good but not great noise figure but in special circumstances it can provide an overall better payload noise figure. These circumstances are that more than 100% bandwidth is required and a power limiter before the LNA would otherwise be required (Micovic et al., 2005). An example of the frequency band of such a LNA is 0.3–4 GHz (Shih et al., 2009). These devices are also made in MMIC since at least 2005 (Micovic et al., 2005).

5.4.2 LNA Linearity

The LNA and the HPA's preamplifier (Chapter 6) are designed to operate in a nearly linear amplification region. The LNA and the preamp do not have explicit IMP specifications but instead have the extent of their nonlinear behavior characterized by one of the two following parameters.

One parameter is the **1-dB compression point**. It is the point at which the actual gain of the LNA is 1 dB lower than the linear gain, as illustrated in Figure 5.5. As the signal power into the element increases, even far below saturation, the element no longer operates in a linear fashion but starts to create significant harmonics along with the intended output signal. Component amplifiers are normally run with input power at least 10 dB below their 1-dB compression point input (Galla, 1989).

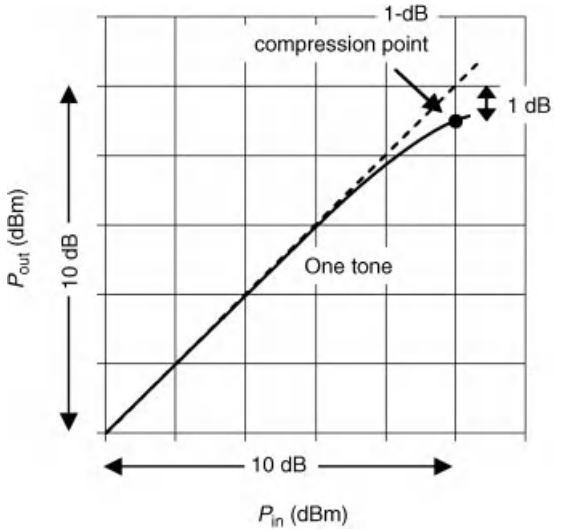


FIGURE 5.5 One-dB compression-point definition.

The other parameter is the **3rd-order intercept point**, whose definition is illustrated in Figure 5.6. Two equal-power tones, whose combined input power is in the linear-amplification region of the amplifier, are input to the amplifier. The tones are at frequencies not far from the middle of the amplifier's passband. The power of one output fundamental and the power of one of the nearby 3rd-order IMPs are measured. The results are sensitive to the fundamental frequencies used, so they must be stated. The amplification curve of one fundamental is extrapolated from the measured point with the slope of 1 dB/dB. The amplification curve of one 3rd-order IMP is extrapolated from its measured point with the slope of 3 dB/dB. The independent axis is the power of one input tone (Agilent, 2000a; Anritsu, 2000). The 3rd-order intercept point is more important than the other orders of intercept point because the 3rd-order IMPs are the largest IMPs.

For an amplifier (but not a frequency converter (Henderson, 1981a)) the 3rd-order intercept-point input is generally 10 dB higher than the 1-dB compression-point input (Henderson, 1990).

5.4.3 LNA Environmental

A LNA without temperature compensation will exhibit higher gain and lower noise figure, the lower the temperature (Microwave Encyclopedia, 2010). In a payload without an active receive phased array, the gain variation can be accommodated by a sufficient dynamic range of the preamp. In a payload with such an array, temperature compensation is necessary. Temperature compensation is achieved either by changing the amplifier bias or by means of a variable attenuator.

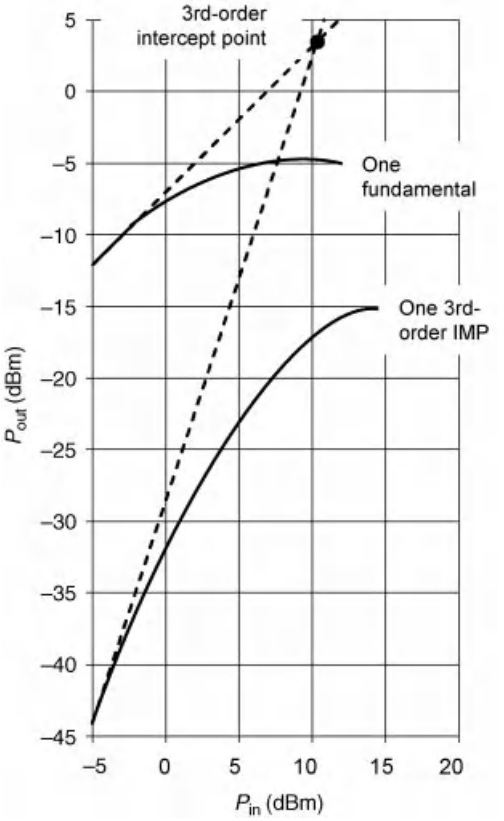


FIGURE 5.6 Third-order intercept-point definition.

An amplifier's gain decreases as the transistor ages (Paine et al., 2000; Chou et al., 2002). This can be compensated by a sufficient dynamic range of the preamp. Also, its noise figure goes up.

5.4.4 LNA Specification

Table 5.1 is an example of the specification parameters for a LNA that is not part of an active receive phased array, that is, where no phase or power matching is required. Examples are given in Table 5.2 of noise figures of state-of-the-art LNAs with return losses no less than 9.5 dB. A return loss occurs when there is a mismatch of impedances at a connection, causing part of a signal reaching the connection in either direction to be reflected. In a specification the assumption is that the connecting element has ideal impedance. Return loss is the ratio of incident power to reflected power and is usually expressed in dB (Sorrentino and Bianchi, 2010).

TABLE 5.1 Example of Parameters in a LNA Specification

Parameter in LNA Specification	Units
Input frequency range	GHz
1-dB compression point input	dBm
Noise figure—EOL over temperature	dB
Gain min and max—BOL over temperature; EOL over temperature	dB
Gain variation—over input frequency range; over any 100 MHz	dB pk-pk
Gain slope	dB/MHz
Gain stability—BOL over temperature; BOL over any 15°C; EOL over temperature	dB pk-pk
Phase linearity for input carrier in some stated power range—over input frequency range; over any 100 MHz	°pk-pk
Return loss—input; output	dB
Operational temperature range	°C

TABLE 5.2 Noise Figures of Some State-of-the-Art LNAs at BOL

Bandwidth Category	Frequency Band	Max NF at 25°C (dB)	Max NF at 60–65°C (dB)	References
Up to 10% bandwidth	6 GHz, 580 MHz wide	1.20	1.45	NEC, 2011
	8 GHz, 500 MHz wide	1.3	1.5	NEC, 2011
	14 GHz, 750 MHz	1.20	1.50	NEC, 2011
	18 GHz, 800 MHz wide	1.70	2.00	NEC, 2011
	27–30 GHz	1.7 (nominal, not max)	2.1 (nominal, not max)	NEC, 2011 (R&D model)
	29 GHz, 1.6 GHz wide	2.15	2.5	Miquel et al., 2006
	44 GHz, 2 GHz wide	3.2	Unknown	Grundbacher et al., 2004
25–75% bandwidth	6–13 GHz	2.5	Unknown	TriQuint, 2005
	21–27 GHz	2.9	Unknown	TriQuint, 2009
Much more than 100% bandwidth	0.3–4 GHz	2.0	Unknown	Shih et al., 2009
	4.5–15 GHz	2.0	Unknown	Micovic et al., 2007

5.5 FREQUENCY CONVERTER

5.5.1 Frequency Conversion Architecture

There are several ways to architect the collection of frequency converters in the payload. Reliability followed by lowest mass are the main drivers of the choice. Figure 5.7 shows some architectures for different situations. Drawing in Figure 5.7a is the simplest architecture, with separate converter units each having its own internal LO. (Filters inside converters are not shown, for simplification at this point in the discussion.) This architecture makes sense when there are only a small number of converters. Sensible redundancy possibilities are the following: the units could each have 2 : 1 redundancy, units could be internally redundant, or there could be redundancy rings if the conversions are all the same. Drawing in Figure 5.7b is for a payload that is a little more complex in regard to the conversions: a medium to large number of converters having the same oscillator frequency, with a master oscillator external to the converters. The master oscillator is so critical that it would have at least 3:1 redundancy (see, e.g., FEI (2006)) and be distributed to the individual converters by a passive network, which is more reliable than an active network. The LO-less converters would have redundancy as in drawing in Figure 5.7a. Drawing in Figure 5.7c is yet another step up. Each converter has its own internal LO which derives from the reference oscillator, and as a collection the LO frequencies are different. Again, because of its criticality the reference oscillator would have at least 3:1 redundancy and the converters would have redundancy as in drawing in Figure 5.7a. Drawing in Figure 5.7d is a variant of Figure 5.7c in which the frequency synthesizers are external to the converters and each capable of creating a large number of different frequencies (Dayaratna et al., 2005).

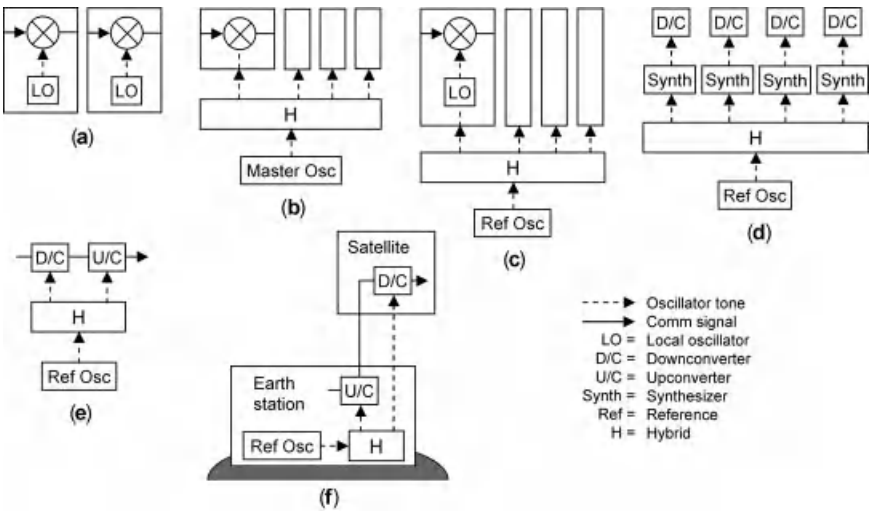


FIGURE 5.7 Some frequency-conversion architectures (redundancy not shown).

With the drawings in the second row of the figure, we start a series of converter architectures with a different theme, that of minimizing phase noise by using the same reference oscillator for both up- and downconversions. We address this in Section 5.3, after we introduce phase noise.

5.5.2 Phase Noise Introduction

Phase noise is an unwanted variation in the signal phase that comes from the communications-system oscillators. Phase noise has a spectrum that goes down to **direct current (DC)**, namely 0 Hz. A spectrum plot has y-axis labeled $\mathcal{L}$, the **single-sideband (SSB) phase noise**. The name is something of a misnomer since $\mathcal{L}$ was redefined more than 20 years ago to have its current definition below:

$$\mathcal{L}(f) = \tfrac{1}{2}\mathcal{S}_\phi(f) \quad \text{where } \mathcal{S}_\phi(f) \text{ is spectrum of phase noise at frequency } f$$

It is, more properly, one side of the double-sided phase noise spectrum. The rms phase noise with frequencies between f_1 and f_2 is then given by

$$\text{Rms phase noise with frequencies between } f_1 \text{ and } f_2 = 2\int_{f_1}^{f_2} \mathcal{L}(f)df$$

A formula for such an integral is given in Appendix 5.A. Usually in an oscillator specification, the frequency f is called the "offset frequency," that is, sideband frequency relative to carrier, harking back to the now-obsolete definition of $\mathcal{L}(f)$. $\mathcal{L}$ is usually given in units of dBc/Hz, meaning power integrated over 1 Hz, stated as relative to the carrier power in dB. The x-axis, frequency, is shown in a logarithmic scale. When plotted in this way, $\mathcal{L}(f)$ has decreasing slope with increasing f.

The spectrum is smooth except for some small discrete components (i.e., sine-waves), also called **spurious phase modulation (PM)**, at low frequencies. However, these frequencies are so low that this spurious PM does not matter—they lie outside the specification band. An example of a spectrum with spurious PM is shown in Figure 5.8. See Section 10.5.2 for a description of what part of the phase noise spectrum should have specifications on it and how phase noise affects the signal.

We return to the converter configurations in the second row of Figure 5.7, which minimize the total phase noise in an upconversion and downconversion by using the same reference oscillator for both conversions. Drawing in Figure 5.7e is for the case where the two conversions occur in the payload and Figure 5.7f for the case where the upconversion occurs in the ground station and the downconversion in the payload. Drawing in Figure 5.7f is what NASA's Tracking and Data Relay Satellite System (TDRSS) does, at least with the satellites F1 through F7, H, I, and J (Lewis, 1996). The advantage comes from the fact that if the reference oscillator's phase is represented by $2\pi f_{\text{ref}}t + \varphi(t)$, where $\varphi(t)$ is the phase noise then the low-frequency phase noise of a LO derived from it with frequency κ times is

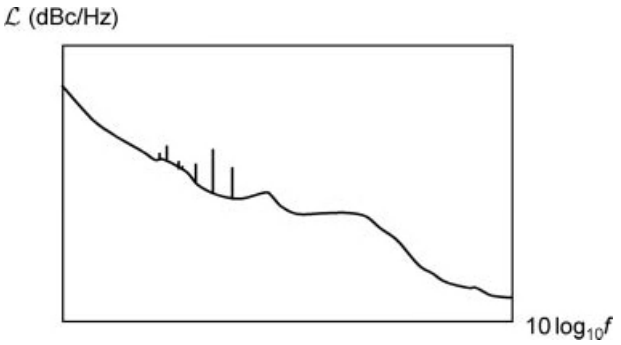


FIGURE 5.8 Example of a phase noise spectrum with spurious phase modulation.

$2\pi\kappa f_{\text{ref}}t + \kappa\varphi(t)$. Suppose that the downconversion LO is κf_{ref} and the upconversion LO is λf_{ref} where $\kappa > \lambda$ and both conversions are low side (defined in the following section). Then the combined phase of the two converters is $2\pi(\kappa - \lambda)f_{\text{ref}}t + (\kappa - \lambda)\varphi(t)$, and the multiplier on the reference oscillator's phase jitter is only $\kappa - \lambda$, indicating a cancellation of most of the low-frequency phase noise from the converters. The power spectral density (psd) of the combined phase, at any given frequency, is $(\kappa - \lambda)^2$ times that of $\varphi(t)$ at the frequency. On the other hand, if unrelated oscillators are used for the two conversions and their respective phase noises are $\kappa\varphi_1(t)$ and $\lambda\varphi_2(t)$, where $\varphi_1(t)$ and $\varphi_2(t)$ have the same psd as $\varphi(t)$, then the psd of the combined phase is $(\kappa^2 + \lambda^2)$ times that of $\varphi(t)$.

5.5.3 Frequency-Converter Unit Architecture and Function

The basic architecture of a frequency converter is shown in Figure 5.9. The architecture is the same for an upconverter as a downconverter. The function can be summarized as follows. The mixer multiplies the RF (IF) signal with the oscillator signal and creates two signals at different frequencies, only one of which is the desired IF (RF) output signal. Suppose $s(t) = \cos(2\pi f_c t + \theta(t))$ is the converter's input signal and $r(t) = \cos(2\pi f_{\text{LO}} t)$ is the LO signal, ignoring signal amplitudes.

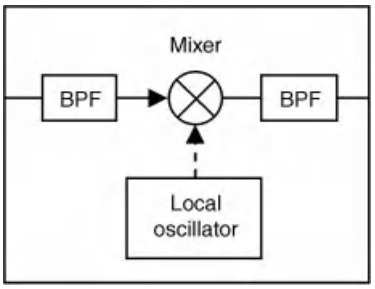


FIGURE 5.9 Frequency-converter basic architecture.

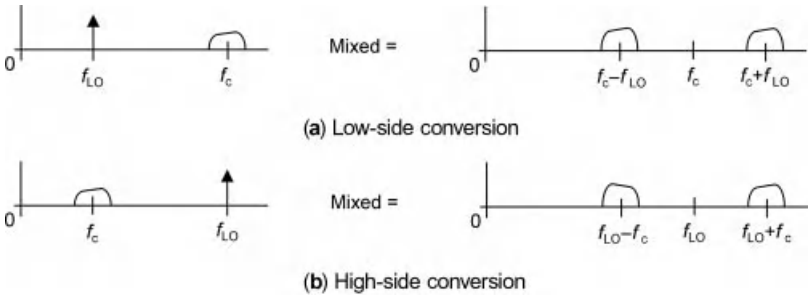


FIGURE 5.10 Low-side and high-side conversions.

Then the mixer output is

$$\begin{aligned} s(t)r(t) &= \cos(2\pi f_c t + \theta(t))\cos(2\pi f_{LO} t) \\ &= \tfrac{1}{2}\cos(2\pi(f_c - f_{LO})t + \theta(t)) + \tfrac{1}{2}\cos(2\pi(f_c + f_{LO})t + \theta(t)) \end{aligned}$$

The frequency-difference term is described as being at the frequency $f_c - f_{LO}$ if that is a positive frequency or at $f_{LO} - f_c$ otherwise. The input and output filters suppress unwanted signals. The LO may be internal to the unit as shown or external. In the following sections, we discuss each element in some detail.

Most of the time $f_{LO} < f_c$, in which case the conversion is a "low-side conversion" as illustrated in Figure 5.10a. The mixer creates two signals, and if the frequency converter is a downconverter the higher-frequency output at $f_c + f_{LO}$ is filtered out, and if it is an upconverter the lower-frequency output at $f_c - f_{LO}$ is filtered out. There are times when a low-side conversion may not be a good idea, for example, when a LO harmonic may lie very near f_c, and then a "high-side conversion" may be the answer. Then $f_c > f_{LO}$. The situation is represented in Figure 5.10b. Notice that the lower-frequency mixer output has been inverted about its output carrier frequency.

There are two input frequencies that can mix to the desired frequency, so the undesired one at the "image frequency" must be removed before the mixing. Figure 5.11 illustrates where the image frequency is for low-side conversions. For both low-side and high-side conversions, in downconversion the image frequency is at $|f_c - 2f_{LO}|$, and in upconversion it is at $f_c + 2f_{LO}$. There are two ways of removing

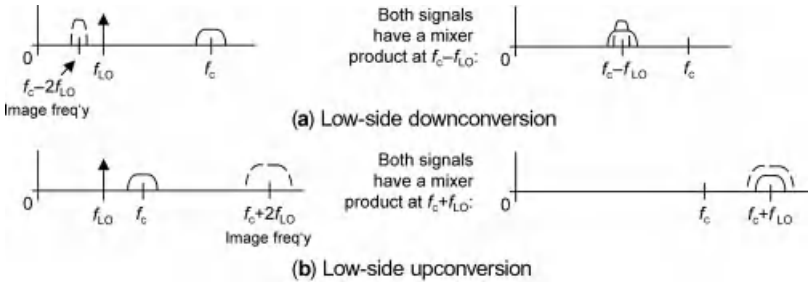


FIGURE 5.11 Examples of image frequency for low-side conversion.

signal at the image frequency: filtering it out with the frequency converter's input BPF or having the mixer do it. Such a mixer is an image-reject mixer (Sorrentino and Bianchi, 2010).

Since mixing is a nonlinear operation, it creates IMPs labeled as follows:

Mixer output (m, n) has frequency $= mf_{\text{in}} - nf_{\text{LO}}$ and order $= |m| + |n|$ where m and n are integers and $m - n \geq 0$

The two output frequencies of an ideal low-side mixer are the mixer products (1,1) and (1,−1), which have order 2. For a downconversion, (1,1) is the desired one and (1,−1) the undesired one, and vice versa for an upconversion. The leaked signals at output are (1,0) and (0,−1). Harmonics are (m,0) and (0,n) for $m > 1$ and $-n > 1$, respectively. The LO harmonics are much stronger than the input-signal harmonics (see following section). The unwanted cross-products are all other allowed (m,n).

5.5.4 Mixer

The mixer itself suppresses unwanted IMPs and leaked signals to a large extent. The possible leakages are illustrated in Figure 5.12. There are four types of mixer circuits: single-ended, single-balanced, double-balanced, and triple-balanced. The single-balanced consists of two single-ended mixers, the double-balanced of two single-balanced, and so on. The most common type is the double-balanced. It has very good LO-to-RF isolation and LO-to-IF isolation. It theoretically creates only one fourth of all intermodulation products, the ones that involve an odd multiple of f_{RF} and an odd multiple of f_{IF} (Henderson, 1981b).

Several aspects of good mixer performance depend on the RF input power being well below the LO power. In most cases, it is at least 20 dB below (Henderson, 1981a).

Interport isolation depends on LO level and temperature. Normally only the LO-to-RF and LO-to-IF isolations are specified because the RF signal level is so much lower than the LO level (Henderson, 1981a).

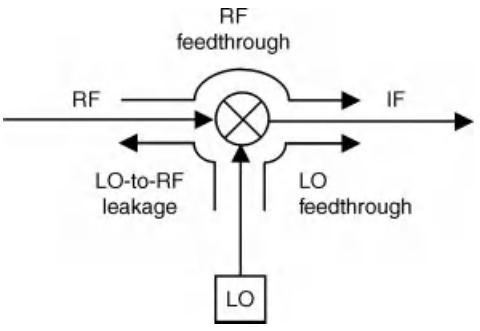


FIGURE 5.12 Mixer isolation definitions (after Agilent (2000b)).

The nonlinearity of the mixer is characterized only in terms of the RF signal level(s), for a given LO level and temperature. Generally the 1-dB compression point input is 5–10 dB lower than the LO input power (Henderson, 1981a). Unlike for an amplifier, the 1-dB compression-point output is generally less than 10 dB below the 3rd-order intercept-point output (Henderson, 1990). The 3rd-order intercept output equals the intercept input plus the mixer conversion gain, which is negative for a passive mixer and positive for an active one (Henderson, 1981a).

5.5.5 Reference Oscillator

Typically in a payload, all the LOs are derived from a single reference oscillator (not counting redundancy in the reference oscillator, see Section 5.5.1). The reference oscillator is most often a quartz **crystal oscillator (XO)** (Fruehauf, 2007). However, Milstar and Adv. EHF (AEHF) have the more expensive but higher-performing **rubidium-vapor atomic oscillator**. Milstar spacecraft have had such an oscillator since 1994 to serve as reference for all its frequency and clock generation, while the AEHF spacecraft use them as reference for the synthesizer (FEI, 2006).

The oscillator is a circuit that takes a voltage signal from a vibrating crystal, amplifies it, and feeds a portion of the amplified signal back to the crystal, which keeps the crystal expanding and contracting at one of its resonant frequencies. A voltage applied across the crystal causes it to contract, and the contraction creates a voltage across the crystal—this is the piezoelectric effect (HP, 1997).

The crystal for a reference oscillator is cut in the form of a plate (Fruehauf, 2007). The cut determines the fundamental frequency and what overtones (harmonic or near-harmonic) and nonharmonic signals the oscillator circuit will generate. A nonharmonic one close to the desired overtone poses a risk because a small environmental change can shift the nonharmonic one on top of the desired overtone, thereby weakening it and creating an "activity dip" (HP, 1997). The thicker the plate, the higher is the order of the overtone that the crystal can vibrate in (Bloch et al., 2002).

The oscillator circuit is subject to environmental influences that change the crystal vibration frequency. The major influence is temperature change. The second most significant influence is time, that is, aging of the crystal itself, its mounting, its electrodes, and so on. The crystal itself ages at a relatively constant rate per decade, from total ionizing dose of radiation. Another influence is the oscillator circuit being turned off and on, which causes an offset in the frequency but long-term does not affect the aging (HP, 1997).

Against temperature variation, one option is temperature-compensation. The **temperature-compensated crystal oscillator (TCXO)** encases the oscillator circuit and temperature-compensating elements in a container. The other temperature option is an oven. The **oven-controlled crystal oscillator (OCXO)** has a heater and heater control in the oscillator circuit, as well as the temperature-sensitive elements in a thermally insulated container. The crystal is kept at its temperature of minimum temperature-sensitivity (HP, 1997).

Aging affects a crystal oscillator by changing its frequency (Fruehauf, 2004). The severity of the radiation effect is minimized when the crystal material is

radiation-hardened by high-temperature heating, when the crystal plate is thick enough to vibrate in its 5th overtone, and when the crystal selected has a monotonically positive aging slope. The latter is good because it offsets the radiation-caused monotonically negative aging slope (Bloch et al., 2002).

A rubidium-vapor atomic oscillator is a rubidium atomic resonator to which a voltage-controlled crystal oscillator (VCXO) is frequency-locked (McClelland et al., 1996). The rubidium atoms emit radiation at particular frequencies corresponding to the energy lost when electrons jump from one energy level to a lower one. The electrons are optically "pumped" to the higher energy states. The VCXO locks to one of the frequencies. The rubidium oscillator's frequency drift due to aging actually slows down to a very small steady-state value. There is no noticeable sensitivity to radiation (Bloch et al., 2002).

5.5.6 Local Oscillator

5.5.6.1 Phase-Locking Introduction LOs are created by **phase-locking** one oscillator to another, so let us introduce phase-locking by looking at the basic **phase-locked loop (PLL)**. It is a feedback circuit that derives a good estimate of a carrier from a noisy input carrier, as shown in Figure 5.13a (Gardner, 1979). The voltage-controlled oscillator (VCO) creates the recovered carrier. The way the PLL works in tracking mode is that the noisy input carrier mixes with the carrier out of the VCO. The mixer functions as a phase detector (the double-frequency term is suppressed) in that its DC output voltage is proportional to the sine of the phase error between the noisy carrier and the recovered carrier. The error voltage then drives the VCO to bring the phase error to zero. Whether the phase error is positive or negative, if it is not too big the error voltage will drive the VCO phase in the right direction. The loop filter smoothes out the voltages that drive the VCO. When the filter is the identity function then the loop is first order (and can track out phase error but not frequency error), and when the filter is an integrator or something like

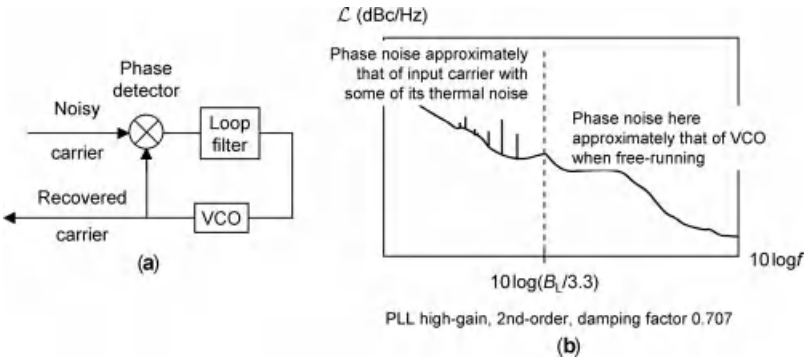


FIGURE 5.13 Carrier recovery: (a) phase-locked loop and (b) phase noise spectrum of recovered carrier.

an integrator then the loop is second order (and can track out both phase and frequency error but not frequency rate error).

The recovered carrier's phase noise is related to the phase noise of both the input carrier and the VCO. A PLL has a **loop bandwidth B_L** determined by the gain in the loop and the loop filter. For a second order, high-gain PLL with damping factor $\zeta = 0.707$, $B_L/3.3$ is the frequency that separates the recovered carrier's phase noise spectrum into two regions, as shown in Figure 5.13b: at frequencies lower than about $B_L/3.3$, the spectrum is that of the input carrier with some of its thermal noise, and at frequencies higher than about that, the spectrum is that of the VCO if it were free-running (Gardner, 1979). Damping factor has a similar meaning as for a spring, coming into play when a transient force is applied and determining how much and for how long it will oscillate before approaching steady state.

5.5.6.2 Dielectric-Resonator Oscillator and Coaxial-Resonator Oscillator The most common way of deriving the LO is to phase-lock a dielectric-resonator oscillator (DRO) to the reference oscillator (Fiedziuszko, 2002). This method has replaced the old method of a simple frequency multiplier because the new method is smaller, less expensive, more reliable, and has lower spurious outputs (Hitch and Holden, 1997). If the LO frequency is below about 2 GHz, instead of a DRO a coaxial-resonator oscillator (CRO) is used (FEI, 2006; AtlanTecRF, 2010; Hitch and Holden, 1997).

The DRO resonates in only a narrow range of frequencies (Silica Valley, 2007). It is a short ceramic cylinder, called a "puck," supported inside a larger conductive enclosure (Trans-Tech, 2007). The puck is coupled to a microstrip line, which is connected to a transistor (Hitch and Holden, 1997). The high-performance ceramics yield DROs with excellent frequency stability over temperature and life and no sensitivity to radiation.

The CRO is a short, dielectric-filled piece of coaxial cable (Hitch and Holden, 1997). The same or similar ceramics are used as in DROs. The line is shorted at one end and a parallel-plate microwave capacitor is used to resonate it. The circuit is connected to a transistor.

The PLL for the LO, shown in Figure 5.14, is somewhat different from the carrier-recovery PLL of Figure 5.13a because now a sampling phase detector replaces the mixer phase detector (Hitch and Holden, 1997). A DRO is shown in Figure 5.14 here but it could just as well be a CRO. The sampling phase detector creates a large

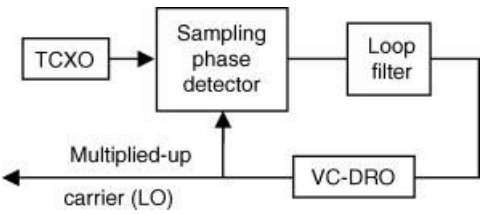


FIGURE 5.14 Dielectric-resonator oscillator phase-locked to reference-oscillator multiple (after Hitch and Holden (1997)).

number of harmonics of the TCXO frequency (Andrews et al., 1990) (up to 170 (Hitch and Holden, 1997)), mixes the harmonics with the DRO's output carrier, and puts out a DC voltage when one of the harmonics matches the DRO's frequency. When there is a DC voltage, it tends to drive the DRO to phase-lock. Therefore, the DRO's free-running (i.e., without voltage-control input) frequency must be close to the TCXO's desired harmonic so that it can be corrected to the harmonic.

The phase noise spectrum of the PLL's output carrier is better than that of the multiplied-up TCXO and better than that of the free-running DRO: it combines the better parts of the two spectra, as discussed in the previous section.

5.5.6.3 Frequency Synthesizer A synthesizer is different from a phase-locked DRO or CRO in that a synthesizer can or does create a large number of different frequencies (Gardner, 1979). A synthesizer phase-locks a VCO to the TCXO or OCXO reference oscillator.

An example of a synthesizer is the 19 identical ones on the JCSAT-5A (called "JCSAT-9" at launch in 2006) satellite (Dayaratna et al., 2005). Each can synthesize over a frequency range 100 MHz wide in 1-MHz steps, and is on-orbit programmable by ground command. The UHF output is upconverted to C-band or L-band by means of a LO locked to the same reference oscillator as the synthesizer. One of the components in the LO is a VCXO phase-locked to a harmonic of the OCXO. This is illustrated in Figure 5.15.

5.5.7 Frequency Converter Linearity

Ignoring the fact that the frequency converter translates the signal's frequency, a nonlinear operation, the linear-operation region of the frequency converter must be specified. The parameters 1-dB compression point or 3rd-order intercept point can be used, as explained in Section 5.4.2, or **carrier-to-3rd-order-IM ratio (C/3IM)** can be used. The latter is measured with two (not counting the LO) equal-power inband tones at some stated power input to the converter. The C/3IM is the ratio at the output of the power in one fundamental to the power in one 3rd-order IMP. The

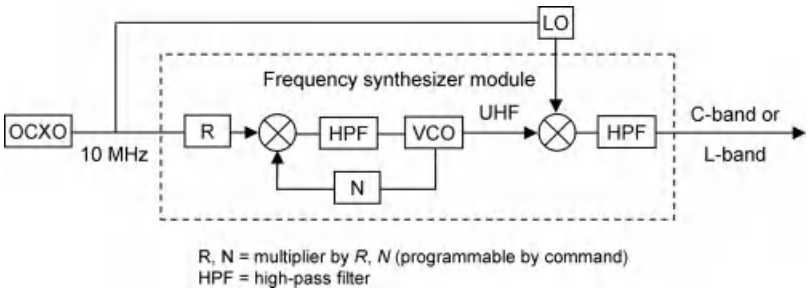


FIGURE 5.15 JCSAT-5A frequency synthesizer (redundancy not shown) (after Dayaratna et al. (2005)).

independent axis is the input power of one carrier (but some people may use the power of both tones together).

5.5.8 Frequency Converter Environmental

Besides the sensitivity of the reference oscillator to the environment, described in Section 5.5.5, the frequency converter itself has sensitivity. Its amplifiers show gain variation. The main sensitivity is to the temperature. If necessary, the frequency converter can incorporate temperature-compensated gain control (e.g., Tramm (2002)). The amplifiers' gains also degrade over life. The preamp needs enough range to accommodate the frequency converter's gain variation.

5.5.9 Frequency Converter Specification

Table 5.3 gives an example of the communications-related parameters in a specification for a space-qualified Ku-band downconverter with a double-balanced mixer (after Narda (2009)). In Table 5.3, items separated by semicolons

TABLE 5.3 Example of Parameters in a Frequency Converter Specification

Parameter in Frequency-Converter Specification	Units
Input frequency range	GHz
Translation frequency	GHz
1-dB compression point input	dBm
Max operational power per input carrier	dBm
Max no. of carriers input	Integer
Noise figure—EOL over temperature	dB
Gain min and max—BOL over temperature; EOL over temperature	dB
Gain variation—over input frequency range; over any 33 MHz wide channel	dB pk-pk
Gain slope	dB/MHz
Gain stability—BOL over temperature; BOL over any 15°C; EOL over temperature	dB pk-pk
Phase linearity for input carrier in some stated power range—over input frequency range; over any 33 MHz wide channel	°pk-pk
Translation frequency stability—over 15°C and 1 day; EOL over temperature including initial settability uncertainty	± ppm
SSB phase noise—upper-bound mask	dBc/Hz
Inband spurious outputs for input carrier of some stated power—each IMP over full input frequency range; other integrated over any 4 kHz wide band	dBc; dBm
Out-of-band spurious outputs for input carrier of some stated power—LO 2nd harmonic; other integrated over any 4 kHz wide band	dBm
Return loss—input; output	dB
Operational temperature range	°C

Source: After Narda (2009).

represent separate parameters. The 15°C temperature range in some of the parameters probably represents a diurnal temperature variation. The gain stability over temperature and life would be the sum of the separate temperature and life terms. Some parameters' numerical limits depend on the power of the input signal(s) used for measuring (Henderson, 1981a). The converter's IMPs and RF and LO feedthrough are together termed "spurious outputs." Another frequency-converter specification could have somewhat different parameters that represent basically the same information.

APPENDIX 5.A

5.A.1 Formula for Integrating Phase Noise Spectrum

Sometimes the payload engineer may want to integrate the payload's measured SSB phase noise spectrum $\mathcal{L}$ defined in Section 5.5.2 over a frequency range by hand. If, for example, the phase noise exceeds its specification at one frequency, he may want to check if nonetheless the spectrum fulfills the unstated specification underlying the specification he was given. The underlying specification, the one that matters, is whether the integral of the phase noise over a certain band is small enough. As stated in Section 5.5.6.1 that band is approximately from $B_L/3.3$ to $R_s/2$, where B_L is the (one-sided baseband) bandwidth in Hz of the carrier-tracking feedback loop in the ground receiver and R_s is the maximum modulation-symbol rate in Hz that the payload will carry. Why this is the band that matters will be explained in Section 10.5.2.

Figure 5.A.1 shows an example of a piecewise-linear approximation to a SSB phase noise spectrum. The engineer can upper bound any measured spectrum in this way, to make it easy to integrate by hand. The phase noise specification itself may also be given in such a way.

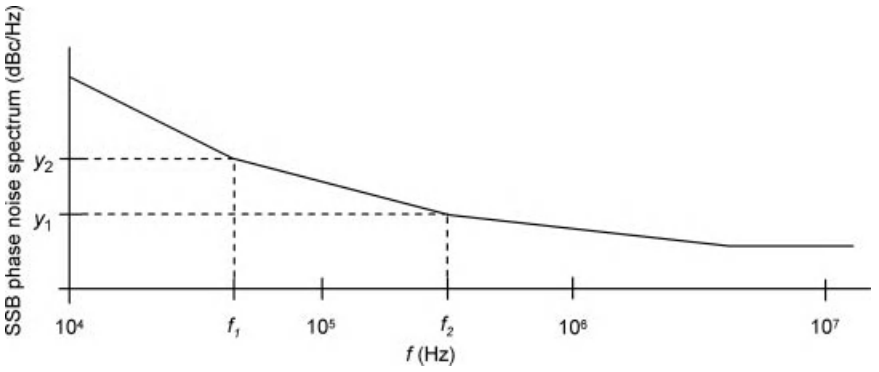


FIGURE 5.A.1 A piecewise-linear approximation to a single-sideband phase noise spectrum.

The rms phase noise contribution that occurs in the interval between frequencies f_1 and f_2 in Figure 5.A.1, on which the spectrum is a straight line, is given by the following, where in this instance $\mathcal{L}(f)$ is not in dBc/Hz but y_i are, $i = 1,2$:

$$\sigma_\phi^2 \text{ contribution between } f_1 \text{ and } f_2 = 2\int_{f_1}^{f_2} \mathcal{L}(f)df$$

$$= \begin{cases} \dfrac{2A}{m+1}(f_2^{m+1} - f_1^{m+1}) & \text{if } m \neq -1 \\ 2A\ \ln(f_2/f_1) & \text{if } m = -1 \end{cases}$$

where

$$A = \frac{10^{y_1/10}}{f_1^m}, \quad m = \frac{y_2 - y_1}{10\log_{10}(f_2/f_1)}, \quad y_i = 10\log_{10}\mathcal{L}(f_i), \quad i = 1,2$$

REFERENCES

Agilent Technologies (2000a). Agilent PSA, Performance spectrum analyzer series, optimizing dynamic range for distortion measurements. Product note. On www.agilent.com. Accessed Jan. 21, 2011.

Agilent Technologies (2000b). Agilent PN 8753-2, RF component measurements—mixer measurements using the 8753B network analyzer. Product note. On www.agilent.com. Accessed Sept. 24, 2010.

Andrews J, Podell A, Mogri J, Karmel C, and Lee K (1990). GaAs MMIC phase locked source. *IEEE Microwave Theory and Techniques Symposium Digest;* 2 (May); 815–818.

Anritsu Co (2000). Intermodulation distortion (IMD) measurements, using the 37300 series vector network analyzer. Application note. On www.anritsu.com/search/en-US. Accessed Jan. 21, 2011.

AtlanTecRF (2010). Stabilised oscillators, CRO & DRO. Product info. On www.atlanticmicrowave.co.uk. Accessed Nov. 23, 2010.

Bloch M, Mancini O, and McClelland T (2002). Performance of rubidium and quartz clocks in space. *Proceedings of IEEE International Frequency Control Symposium;* May 29–31; 505–509.

Briskman RD and Prevaux RJ (2001). S-DARS broadcast from inclined, elliptical orbits. *International Astronautical Congress;* Oct. 1–5; 503–518.

Chou YC, Leung D, Lai R, Grundbacher R, Eng D, Scarpulla J, Barsky M, Liu PH, Biedenbender M, Oki A, and Streit D (2002). Evolution of DC and RF degradation induced by high-temperature accelerated lifetest of pseudomorphic GaAs and InGaAs/InAlAs/InP HEMT MMICs. *IEEE International Reliability Physics Symposium;* Apr. 7–11; 241–247.

Chou YC, Barsky M, Grundbacher R, Lai R, Leung D, Bonnin R, Akbany S, Tsui S, Kan Q, Eng D, and Oki A (2003). On the development of automatic assembly line for InP HEMT MMICs. *International Conference on InP and Related Materials;* May 12–16; 476–479.

Dayaratna L, Ramos LG, Hirokawa M, and Valenti S (2005). On orbit programmable frequency generation system for JCSAT 9 spacecraft. *IEEE Microwave Theory and Techniques Symposium Digest;* June 12–17; 1191–1194.

Fiedziuszko SJ (2002). Satellites and microwaves. *International Conference on Microwaves, Radar and Wireless Communications;* 3 (May 20–22); 937–953.

Frequency Electronics, Inc (2006). Space—FEI capabilities. July. On www.freqelec.com/tech_lit.html. Accessed Sept. 28, 2010.

Fruehauf H, Frequency Electronics, Inc (2004). Technical literature: understanding oscillator specs. Presentation. Aug. On www.freqelec.com/tech_lit.html. Accessed Sept. 28, 2010.

Fruehauf H, Frequency Electronics, Inc (2007). Technical literature: precision oscillator overview. Presentation. Apr. On www.freqelec.com/tech_lit.html. Accessed Sept. 28, 2010.

Galla TJ (1989). TriQuint Semiconductor Technical Library: cascaded amplifiers. WJ Technote. On www.triquint.com/prodserv/tech_info/WJ_tech_publications.cfm. Accessed Sept. 28, 2010.

Gardner F (1979). *Phaselock Techniques*, 2nd ed. New York: John Wiley & Sons, Inc.

Grundbacher R, Chou Y-C, Lai R, Ip K, Kam S, Barsky M, Hayashibara G, Leung D, Eng D, Tsai R, Nishimoto M, Block T, Liu P-H, and Oki A (2004). High performance and high reliability InP HEMT low noise amplifiers for phased-array applications. *IEEE MTT-S International Microwave Symposium Digest;* 1 (June 6–11); 157–160.

Henderson BC (1981a). TriQuint Semiconductor Technical Library, WJ Technical Publications: mixers: part 1, characteristics and performance. WJ Tech-note. Revised 2001. On www.triquint.com/prodserv/tech_info/WJ_tech_publications.cfm. Accessed Sept. 28, 2010.

Henderson BC (1981b). TriQuint Semiconductor Technical Library: mixers: part 2, theory and technology. WJ Tech-note. Revised 2001. On www.triquint.com/prodserv/tech_info/WJ_tech_publications.cfm. Accessed Sept. 28, 2010.

Henderson BC (1990). TriQuint Semiconductor Technical Library: mixers in microwave systems (part 2). WJ Tech-note. Revised 2001. On www.triquint.com/prodserv/tech_info/WJ_tech_publications.cfm. Accessed Sept. 28, 2010.

Hewlett Packard (1997). Fundamentals of quartz oscillators. Application note 200-2. On www.hpmemory.org/ressources/resrc_an_01.htm. Accessed Nov. 24, 2010.

Hitch B and Holden T (1997). Phase locked DRO/CRO for space use. *Proceedings of IEEE International Frequency Control Symposium;* May 28–30; 1015–1023.

Lewis T (1996). TDRSS technical information package for 2nd TDRSS Workshop 25–26 June 1996. Sponsored by NASA HQ's Office of Space Communications and Goddard Space Flight Center. On msp.gsfc.nasa.gov/TUBE/techinfo.htm. Accessed Nov. 22, 2010.

McClelland T, Bhaskar N, and Mallette LA (1996). Rubidium locked oscillator life testing. *Proceedings of IEEE Aerospace Applications Conference;* 4 (Feb. 3–10); 469–478.

Micovic M, Kurdoghlian A, Moyer HP, Hashimoto P, Schmitz A, Milosavljevic I, Willadsen PJ, Wong W-S, Duvall J, Hu M, Wetzel M, and Chow DH (2005). GaN MMIC technology for microwave and millimeter-wave applications. *IEEE Compound Semiconductor Integrated Circuit Symposium;* Oct. 30–Nov. 2; 173–176.

Micovic M, Kurdoghlian A, Lee T, Hiramoto RO, Hashimoto P, Schmitz A, Milosavljevic I, Willadsen PJ, Wong W-S, Antcliffe M, Wetzel M, Hu M, Delaney MJ, and Chow DH

(2007). Robust broadband (4 GHz–16 GHz) GaN MMIC LNA. *IEEE Compound Semiconductor IC Symposium;* Oct. 14–17; 1–4.

Microwave Encyclopedia (2006). Microwave FET tutorial. Jan. 22. On www.microwaves101.com. Accessed Dec. 7, 2010.

Microwave Encyclopedia (2010). Power amplifiers. Oct. 15. On www.microwaves101.com. Accessed Jul. 1, 2011.

Miquel C, Cayrou JC, and Cazaux JL (2006). Flexible Ka-band low noise amplifier subsystem for oncoming satellite payloads. *Proceedings of European Microwave Conf*erence; Sept.; 890–893.

Narda Microwave West (2009). Space qualified converters. Brochure. On www.rfglobalnet.com. Accessed Jan. 6, 2009.

NEC Toshiba Space Systems Ltd (2011). Low-noise amplifier product data sheets. On www.nec.com/global/solutions/space/satellite_communications/index.html. Accessed Jan. 27, 2011.

Paine B, Wong R, Schmitz A, Walden R, Nguyen L, Delaney M, and Hum K (2000). Ka-band InP HEMT MMIC reliability. *Proceedings of GaAs Reliability Workshop;* Nov. 5; 21–44.

Shih SE, Deal WR, Yamachi D, Sutton WE, Chen YC, Smorchkova I, Heying B, Wojtowicz M, and Siddiqui M (2009). Design and analysis of ultra wideband GaN dual-gate HEMT low noise amplifiers. *IEEE MTT-S Microwave Symposium Digest*; June 7–12; 669–672.

Silica Valley.com (2007). Dielectric resonator. On www.electronics-manufacturers.com/info/rf/microwave-electronics. Accessed Sept. 28, 2010.

Sorrentino R and Bianchi G (2010). *Microwave and RF Engineering*. Chichester, UK: John Wiley & Sons, Ltd.

Sowers JJ, Willis M, Tieu T, Findley W, and Hubbard K (2001). A space-qualified, hermetically-sealed, Ka-band LNA with 2.0 dB noise figure. *Technical Digest, IEEE GaAs IC Symposium;* Oct. 21–24; 156–161.

Tramm FC (2002). Compact frequency converters for a Ka-band telecommunications satellite payload. *AIAA International Communication Satellite Systems Conference;* May 12–15. On www.as.northropgrumman.com/products/tech_publications/pdfs. Accessed Oct. 4, 2010.

Trans-Tech, Inc (2007). Introduction and applications for temperature stable dielectric resonators. Application note. On www.trans-techinc.com/list_documents.asp?doctype=38. Accessed Nov. 24, 2010.

TriQuint Semiconductor (2005). 6–13 GHz low noise amplifier TGA8399-SCC. Product data sheet; Oct. 14. On www.triquint.com/prodserv/markets/military/space.cfm. Accessed Jan. 27, 2010.

TriQuint, Semiconductor (2009). K band packaged low noise amplifier TGA4506-SM. Product data sheet, rev A; Sep. On www.triquint.com/prodserv/markets/military/space.cfm. Accessed Jan. 27, 2010.

Wikipedia. On en.wikipedia.org/wiki/Main_Page.

Yeung TK, Gregg H, and Morgan I (1993). Lightweight low noise space qualified L-band LNA/filter assemblies. *European Conference on Satellite Communications;* Nov. 2–4; 122–127.

CHAPTER 6

PREAMPLIFIER AND HIGH-POWER AMPLIFIER

6.1 INTRODUCTION

The high-power amplifier (HPA) provides the RF power for a payload downlink. Before the signal goes to the HPA, the preamplifier boosts the signal to a level proper for input to the HPA. We may call these two together the **HPA subsystem**. There are two types of HPA subsystem: the **traveling-wave tube amplifier (TWTA)** subsystem and the **solid-state power amplifier (SSPA)**. A subsystem named for the TWTA is not normally identified, but for the purposes of this book it must be so. The TWTA subsystem is more common than the SSPA. One of the leading global satellite manufacturers, which makes its own TWTA subsystems and SSPAs, reports that in the periods 1980–2004 and 2004–2008, the number of its TWTAs on orbit on commercial payloads was twice the number of SSPAs (Weekley and Mangus, 2004; Nicol et al., 2008).

The HPA subsystem has the following functions:

- Channel **preamplification** for the HPA, with the flexibility to make the downlink power independent of the uplink power over a wide range of uplink power
- Predistortion (optional) to counteract the HPA's nonlinear amplification characteristics
- High-power amplification
- Power supply for rest of the functions.

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

The rest of this chapter addresses the following topics:

- *Section 6.2*: HPA concepts and terms different from those used for other units: the HPA's nonlinearity description and specification, as well as power efficiency
- *Section 6.3*: Summary of the differences and similarities in performance of the TWTA subsystem and the SSPA
- *Section 6.4*: TWTA subsystem: the subsystem architecture, subsystems in the payload, subsystem elements' architecture and technology, a list of specification parameters with current performance of the key parameters, a subsystem with unusual flexibility, and environmental considerations
- *Section 6.5*: SSPA unit: similar topics as for the TWTA subsystem.

6.2 HIGH-POWER AMPLIFIER CONCEPTS AND TERMS

6.2.1 HPA Nonlinearity Description

Both types of HPA amplify in a significantly nonlinear way. The SSPA and the typical-bandwidth TWTA are said to provide a **bandpass nonlinearity**, which we define. Suppose that the RF signal (including noise, interference, and signal distortions) into the amplifier is represented as follows:

$$\text{Amplifier input is } \sqrt{2P_{\text{in}}(t)}\cos\left(2\pi f_c t + \theta(t)\right) \text{ where } f_c \text{ is the carrier frequency}$$

$P_{\text{in}}(t)$ is the **instantaneous (rms) power** of the input signal at time t, equal to the signal's square magnitude averaged over a carrier cycle. We now drop the t reference in order to make the amplifier functioning more transparent. As a function of P_{in}, the amplifier produces a power P_{out} and shifts the signal's phase by ϕ:

$$\text{Amplifier output is } \sqrt{2P_{\text{out}}}\cos\left(2\pi f_c t + \theta + \phi\right) \text{ where } P_{\text{out}} \text{ and } \phi \text{ are functions of only } P_{\text{in}}$$

The $\boldsymbol{P_{\text{out}}}$ **versus** $\boldsymbol{P_{\text{in}}}$ **curve** and the **phase shift versus** $\boldsymbol{P_{\text{in}}}$ **curve** completely define the bandpass nonlinearity. These curves are by definition frequency-independent. The case where the frequency dependence is significant, for very-wideband (15% bandwidth or more) TWTAs, is characterized in an augmented way (Section 13.3.6).

The HPA cannot put out an indefinitely large amount of power, so at some point its gain starts to decrease as P_{in} increases. This is gain compression and was illustrated in Section 5.4.2. The range of P_{in} values that are so small there is no (significant) gain compression, that is, the amplification is very nearly linear, is the **small-signal** region of amplification. The HPA **saturates** when it puts out its maximum RF output power. The input power at this point is $\boldsymbol{P_{\text{in sat}}}$ and the output power is $\boldsymbol{P_{\text{out sat}}}$. When P_{in}

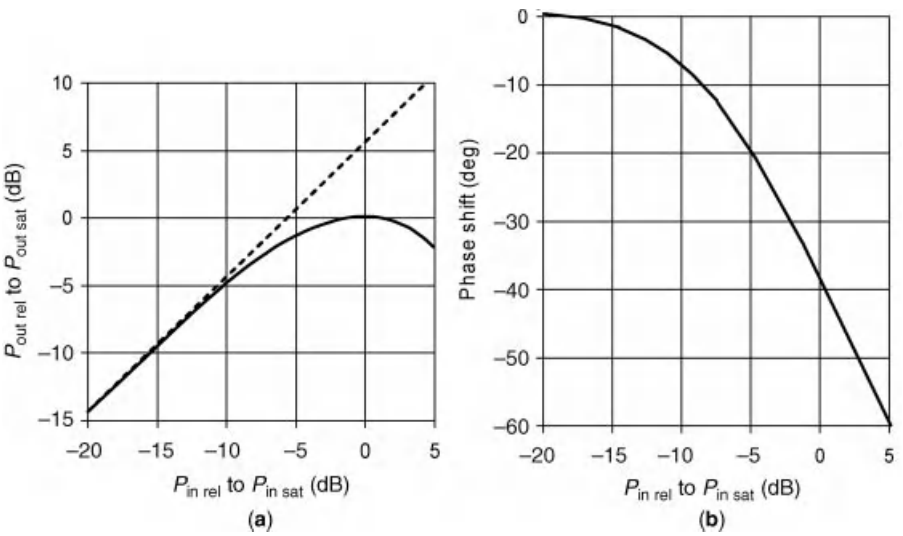


FIGURE 6.1 Example of TWTA's (a) P_{out} versus P_{in} curve and (b) phase shift versus P_{in} curve.

is below $P_{in\ sat}$, the HPA is said to be **backed off**. **Input backoff (IBO)** is the ratio $P_{in\ sat}/P_{in}$ in dB. When P_{in} exceeds $P_{in\ sat}$, the HPA is said to be **in overdrive** or **overdriven**. **Output backoff (OBO)** is the ratio $P_{out\ sat}/P_{in}$ in dB and is positive.

Figure 6.1 shows an example of the P_{out} versus P_{in} and phase shift versus P_{in} curves for a typical TWTA with **continuous-wave (CW)** (= tone) input. As is usual, both axes of the P_{out} versus P_{in} curve are logarithmic. For the phase shift versus P_{in} curve, the y-axis is in degrees while the x-axis is logarithmic. The phase shift is normalized, as always, to be 0° at small signal. In this example, powers are shown relative to saturation power. (Sometimes IBO and OBO are used instead for the axes but with the opposite sign from their formal definition so that increasing power goes in the normal directions.) Note that the gain starts to compress at P_{in} well below $P_{in\ sat}$, a characteristic of a TWTA. At saturation the gain compression is about 6 dB and the phase shift is about −40°.

A SSPA has different behavior from a TWTA near saturation. An example of SSPA nonlinear behavior is shown in Figure 6.2, for the Inmarsat-3 L/C-band upconverter/SSPA unit (Khilla and Leucht, 1996). The C-band SSPA is used on the fixed-terminal links. The upconverter is operating in its linear region, so the nonlinear behavior is that of the SSPA. The P_{out} versus P_{in} curve is nearly linear for P_{in} up to within just a few dB of the 1.6-dB compression point at the right edge of the graph. The SSPA's phase shift at this point is much smaller in magnitude than a TWTA's. The 2-dB compression point **P2dB** is referred to almost exclusively for a SSPA rather than the saturation point. The reason is that P2dB should not be exceeded by much, because otherwise the GaAs transistor's drain will suffer breakdown and the transistor will be damaged (Khilla, 2011d). Sometimes in the

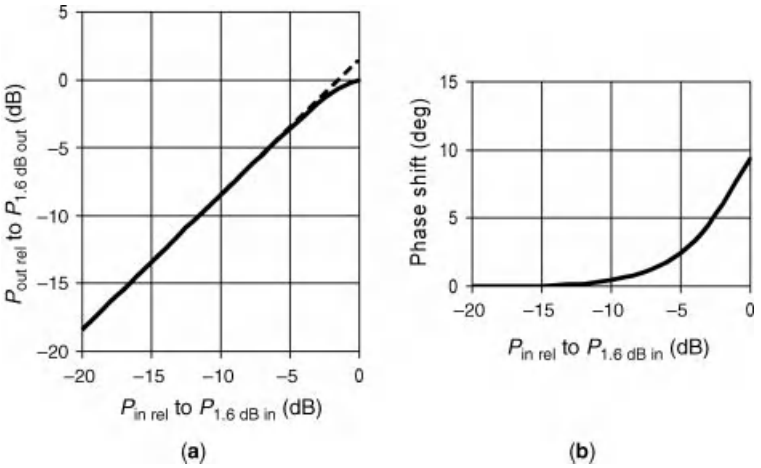


FIGURE 6.2 Example of SSPA's (a) P_{out} versus P_{in} curve and (b) phase shift versus P_{in} curve (after Khilla and Leucht (1996)).

literature P2dB or thereabouts is called the "saturation point." Figure 6.3 shows the over-driven behavior of a small transistor, which is what a SSPA's would be if it were overdriven (Khilla, 2011d). True saturation is reached, with P_{out} about 1 dB greater at saturation than at P2dB. The P_{out} versus P_{in} curve does not turn down as it does for a TWTA. The phase shift (not shown) keeps growing beyond P2dB, approximately linearly with the P_{in} logarithm.

The HPA's saturation point depends on the particular kind of signal input. $P_{out\ sat}$ is the biggest for a noiseless CW; such $P_{out\ sat}$ is the most frequently quoted number for HPA performance. For any other kind of signal input, not only is $P_{out\ sat}$ smaller but $P_{in\ sat}$ is also different. This is illustrated in the following section.

The HPA's **operating point P_{op}** corresponds to the P_{in} equal to the long-term average (rms) power of the operational input RF signal, where the average is taken over a time much, much longer than the inverse of the signal's noise bandwidth. If, somehow, only one tone could be amplified, the instantaneous rms power would

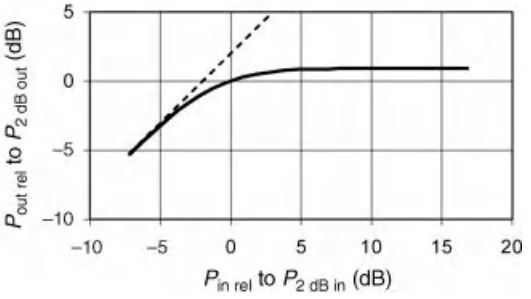


FIGURE 6.3 Example of small transistor's P_{out} versus P_{in} curve (after Khilla (2011d)).

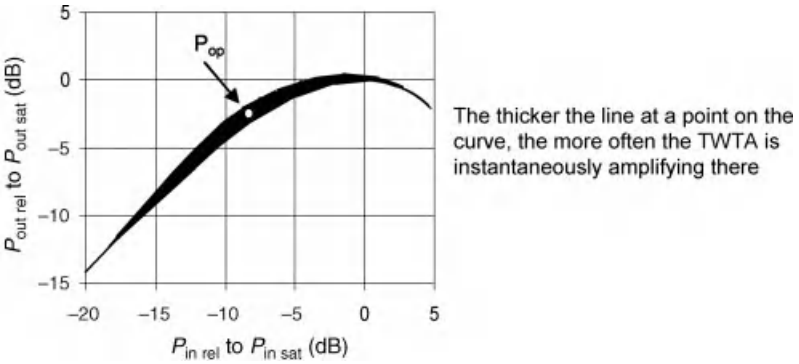


FIGURE 6.4 Example of how instantaneous P_{in} varies around operating-point, for realistic signal input to TWTA.

always equal the long-term average power, but otherwise the instantaneous rms power varies in time about P_{in} of the operating point. Even if the signal is nominally a CW, it will have noise on it; if the signal is modulated, it will additionally have instantaneous level variations. So the instantaneous P_{in} will run up and down relative to the operating-point P_{in}. Even if the operating point is a few dB below saturation, there will be instants in which the HPA is in overdrive. Conversely, even if the operating point is at saturation, there will be instants in which the HPA is in its small-signal region. This is illustrated in Figure 6.4.

The intermodulation products (IMPs) that the HPA creates are mostly inband or near-out-of-band. Since the inband ones cannot be removed, they must be low enough not to be a problem. IMPs are discussed extensively in Section 8.4.2.

6.2.2 HPA Nonlinearity Specification Parameters

There are four common ways in which the HPA nonlinearity is specified (but see caveats after the list):

- P_{out} versus P_{in} curve and phase shift ϕ versus P_{in} curve. Usually each curve is specified by a **mask**, that is, a curve made of straight segments forming the upper bound.
- Derivative of P_{out} versus P_{in} curve and derivative of phase shift versus P_{in} curve, the **AM/AM conversion** and **AM/PM conversion** curves, respectively (see, e.g., Agilent (2000)). (The terms are sometimes used for other things.) The units of the former are dB/dB and of the latter deg/dB. AM/AM conversion equals 1 at small signal and 0 at saturation, and AM/PM conversion equals 0 at small signal. Usually just maximum AM/AM conversion and maximum absolute value of AM/PM conversion are specified, not the entire curves.
- ***C/3IM***: The curve of combined power in the two output fundamentals relative to combined power in the two nearest 3rd-order IMPs, versus combined power

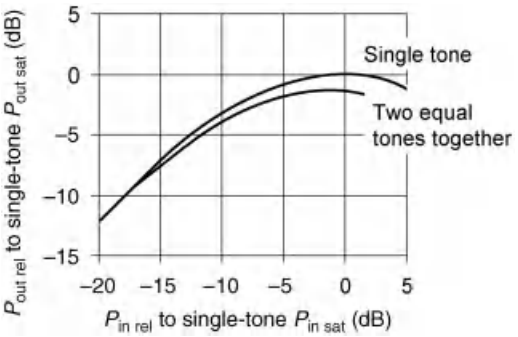


FIGURE 6.5 Different saturation points for one-tone and two-tone TWTA input.

of the two equal-power input tones. (Sometimes a slightly different definition is used.) Input power is in dBm and the output power ratio is in dB. The HPA's $P_{out\ sat}$ for two-tone input is lower than it is for one-tone input, and $P_{in\ sat}$ is lower. A typical example for a TWTA is illustrated in Figure 6.5 (Khilla, 2011c). The C/3IM curve may depend on the particular tone frequencies used, for a wide band.

- ***NPR (Noise-to-Power Ratio)***: The power ratio of all the IMPs in one narrow band inside the signal band, to the signal power in the same narrow band, at HPA output. It is normally measured by applying a large number of equal-power, evenly spaced tones to the HPA input, but leaving one out. The worst-case ratio over the signal band is reported, which is usually to be found mid-band. The definition is illustrated in Figure 6.6.

Knowing both the AM/AM conversion curve and AM/PM conversion curve is conceptually equivalent to knowing both the P_{out} versus P_{in} and ϕ versus P_{in} curves, by integration of the former set. The integrated AM/AM conversion curve could be shifted to its proper level from knowledge of the small-signal gain, which is found by other means. The integrated AM/PM conversion curve could be shifted properly because the phase shift at small signal is taken to be 0°. However, the accumulated errors in the integrations are large, so in fact the conversion curves do not provide the same information. The conversion curves are also inadequate in the case when TWTAs must be matched in their absolute phase delay.

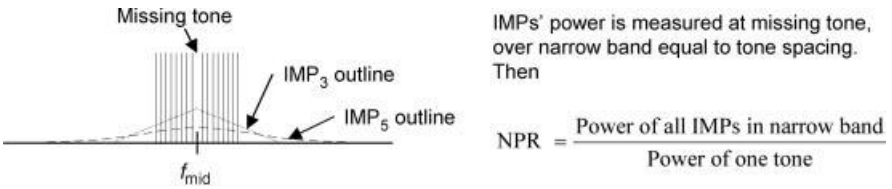


FIGURE 6.6 Noise–power ratio definition.

The C/3IM and NPR specifications are each useful when the operational signals into the HPA will be very much like the test signals, but they are otherwise not very useful (see Section 8.4.1 for detailed discussion on this, based on elementary probability theory).

6.2.3 Power Efficiency

The HPAs are the biggest consumer of satellite DC power for a nonprocessing payload. An operating point at saturation is the most power-efficient but is unsuitable for a signal with much amplitude variation. However, the farther away from saturation, the lower the power efficiency.

An important performance parameter of a HPA is its **power-added efficiency (PAE)**, equal to $(P_{\text{RFout}} - P_{\text{RFin}})/P_{\text{DC}}$ where P_{DC} is the DC power required. The TWTA's PAE is the product of the TWT's PAE with the EPC's power efficiency, and similarly for a SSPA's HPA section. The EPC's power efficiency equals the DC power that the EPC supplies to the HPA divided by the EPC's DC power draw.

For the entire HPA subsystem, the relevant parameter is simply **power efficiency** because the input RF power is negligible. For a TWTA subsystem the power efficiency is well approximated by $\text{PAE}/(1 + (P_{\text{DC LCAMP}}/P_{\text{DC TWTA}}))$. A (L)CAMP uses only 1–4 W of DC power. An example of power efficiency calculation is the following. For a TWTA with a PAE of 65% and output power of 100 W for a saturating CW and for a LCAMP that consumes 2 W, the TWTA subsystem's power efficiency would be 64%. If the TWTA's power efficiency is known then the subsystem's power efficiency is $(\text{TWTA power efficiency})/(1 + (P_{\text{DC LCAMP}}/P_{\text{DC TWTA}}))$.

In product literature for a TWTA, the PAE is typically stated only at CW saturation. For a SSPA the power efficiency is stated at the CW saturation point and/or at a backoff at which NPR is about 15 dB. For a true efficiency comparison, the TWTA PAE and the SSPA efficiency must be given by the manufacturers at the intended operating point.

6.3 TRAVELING-WAVE TUBE AMPLIFIER VERSUS SOLID-STATE POWER AMPLIFIER

Comparing TWTAs and SSPAs is much like comparing apples and oranges. An unlinearized SSPA and a LTWTA have about equally good overall nonlinearity characteristics (Bosch et al., 2003), but a linearized SSPA has better. However, the power efficiency of a TWTA subsystem for the same NPR is better than that of a SSPA. Moreover, TWTAs are more powerful, and so on. Most of the time, the choice is clear.

TWTAs and SSPAs are used in different frequency bands, generally, although both are used at C-band. At L- and S-bands, SSPAs are more often used because of their size (dimensions) advantage (Kaliski, 2009). The advantage is great for active phased-array applications, where the SSPA is located in the limited space right behind its corresponding radiating element. As Kaliski (2009) points out, an S-band

exception is the TWTA-using Sirius Radio satellites. Because the satellite has to put out nearly 4000 W of RF power, TWTAs have to be used for their high PAE (Briskman and Prevaux, 2001). At Ku-band and above almost exclusively TWTAs are used for their high PAE (Kaliski, 2009).

C-band with an output power of 20–40 W is about the only case where TWTAs and SSPAs can compete head-on (Bosch et al., 2003). A spacecraft-level tradeoff study showed even at 60 W no overall mass advantage of the SSPA compared to the LTWTA. When the supporting hardware (heat sink, heat pipe, battery, solar array, etc.) mass was included with the HPA mass, each SSPA was found to need 2 kg more mass than each LTWTA (Bosch et al., 2003).

Comparative reliability of TWTAs and SSPAs is no longer a concern. For a long time, it was thought that SSPAs were more reliable than TWTAs, but since 2004 it has been generally known that this is not the case, from reports by a leading satellite manufacturer which makes its own SSPAs and TWTAs (Weekley and Mangus, 2004). It was also long thought that the SSPA had the advantage of often degrading "gracefully," while the TWTA failed catastrophically. This was also shown not to be true in any significant manner; they both fail catastrophically (Weekley and Mangus, 2005). The 2004 result was based on 800 SSPAs and 1700 TWTAs in orbit; the 2008 result by the same team was based on 950 SSPAs and 2000 TWTAs (Nicol et al., 2008). In 2008, the running failure rates of SSPAs and TWTAs were close, with the TWTA rate being a little lower. The failure rate for both SSPAs and TWTAs had decreased from the year 2001 to 2008. The newer SSPAs were showing a lower failure rate than earlier ones, which brought the failure rate of recently launched SSPAs and TWTAs to virtually identical (Nicol et al., 2008).

TABLE 6.1 Summary of Commercial TWTA Subsystem and SSPA Comparison

Characteristic	TWTA Subsystem Property	SSPA Property
Frequency range	L-band to Ka-band	L-band to C-band
Bandwidth	4–10%	2–9%
P_{out} and power efficiency for CW	275 W and 62% at S-band 125 W and 65% at C-band 130 W and 61% at K-band	40 W and 55% at S-band 60 W and 47% at C-band
P_{out} and power efficiency for multiple carriers	—	20 W and 35% at S-band
Antenna application	Reflector antennas	Active and semiactive phased arrays (either directly radiating or as feeds for reflector antenna)
Size	Helical section of TWT has length approximately proportional to wavelength, for given gain	Smaller than TWTA at C-band and lower frequencies
Cooling method	Conduction or direct radiation	Conduction

It has been pointed out (Phelps, 2008) that TWTA reliability has improved much faster than the military-standard electronic-parts reliability handbook (MIL-HDBK-217, 1995) gets updated. The same must be true of the SSPA. This handbook is now to be taken for reference only but in the past was used to derive specifications. The actual reliability of a 50 W C-band TWT was found to be 20 times better than the number given in the handbook (Phelps, 2008).

Table 6.1 gives a summary of performance differences between TWTA subsystems and SSPAs on commercial satellites. In Sections 6.4.7 and 6.5.6 are detailed examples of current TWTA and SSPA P_{out} and power-efficiency performance, respectively. TWTA cooling is addressed in Section 6.4.9.

6.4 TRAVELING-WAVE TUBE AMPLIFIER SUBSYSTEM

6.4.1 Introduction

Figure 6.7 shows a typical breakdown of a TWTA subsystem into units. We must know the typical breakdown before we can look into the payload-level issues in Section 6.4.2. In Section 6.4.3 we look at the variations in subsystem architecture. The units, terminology, and functions in a typical breakdown are as follows:

- ***Channel Amplifier (CAMP)***: The preamplifier unit. When it incorporates the linearizer function, it is the **linearizer-CAMP (LCAMP)** unit. When referring to the CAMP when it does not matter whether or not it contains a linearizer function, we write **(L)CAMP**. The CAMP has a wide range of gain, for example, 30 dB, which enables it to take almost any level of signal input to any desired TWTA-input level. (In the signal path before the CAMP, there are only fixed levels of amplification.) The CAMP has two operation modes, **fixed-gain mode (FGM)** and **automatic level-control (ALC)** mode. In FGM, it provides the commanded amount of gain, no matter what the input signal level is as long as it is not too high. In ALC mode, it provides the commanded level of output signal, no matter what the input signal level is as long as it is within the capability range. In all cases the CAMP puts out very much less than 1 W.
- ***Linearizer*** *(Optional)*: Sometimes a unit itself but more often part of the CAMP. A nonlinear device to predistort the signal inversely to the TWTA's

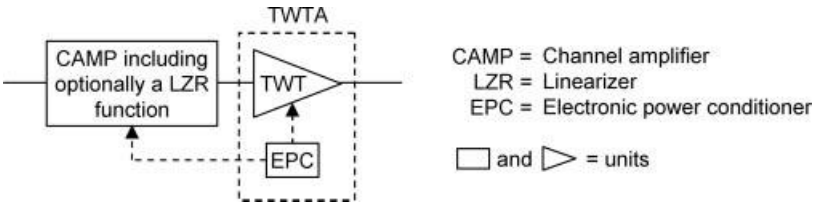


FIGURE 6.7 Typical breakdown of TWTA subsystem into units.

nonlinearity in gain and phase shift. A linearized TWTA can be operated closer to saturation and thus with higher PAE.

- ***Traveling-Wave Tube***: The TWT and EPC units (see immediately below) together form the TWTA, the HPA in this subsystem. A **linearized TWTA (LTWTA)** is a TWTA together with a linearizer function. A **(L)TWTA** is a TWTA irrespective of whether or not there is a linearizer in the subsystem. For TWTAs on commercial payloads today, the TWTA's gain ranges from 35 to 60 dB (Thales, 2009), and its saturated CW P_{out} ranges from 10 to 300 W (L-3, 2009a; Thales, 2009).
- ***Electronic Power Conditioner (EPC)***: The unit that provides DC power to the TWT and the rest of the subsystem.

The rest of this section on the TWTA subsystem is organized differently from other such sections on units, for example, the LNA, since the TWTA subsystem must in some regards be treated as a whole and in other regards as separate units:

- *Section 6.4.3*: TWTA subsystems in payload, traditional and flexible
- *Section 6.4.2*: Subsystem architecture: embodiments of the TWTA subsystem
- *Section 6.4.4*: CAMP, including communications-related specification parameters
- *Section 6.4.5*: Linearizer, including LTWTA nonlinear performance
- *Section 6.4.6*: TWTA, including (L)TWTA specification parameters
- *Section 6.4.7*: Current subsystem performance in key parameters
- *Section 6.4.8*: Flexible subsystem
- *Section 6.4.9*: Subsystem environmental.

6.4.2 TWTA Subsystems in Payload

6.4.2.1 Architecture in Payload, Traditional and Flexible In traditional payload architecture, typically the active TWTAs amplify separate sets of signal(s), that is, each signal is amplified by only one TWTA. This is the most common architecture.

A relatively new, flexible architecture is the **multiport amplifier (MPA)** (Egami and Kawai, 1987). It can handle unbalanced traffic among channels and traffic variation over time. A 4×4 MPA is shown in Figure 6.8. The MPA divides all the input signals equally, then each of multiple TWTAs amplifies the same mix of signals with the same gain, and finally the MPA recombines the signals so as to reconstruct the separate signals. It is possible for one signal to receive all the amplification, several signals to divide it equally, or any condition in between. A limitation is that all the signals must be in different frequency channels. The dividing and combining are accomplished by the input and output "Butler matrices," respectively, which together form a nearly lossless multiport network. The input and output matrices can be either identical or the reverse of each other. The TWTAs must closely **match**

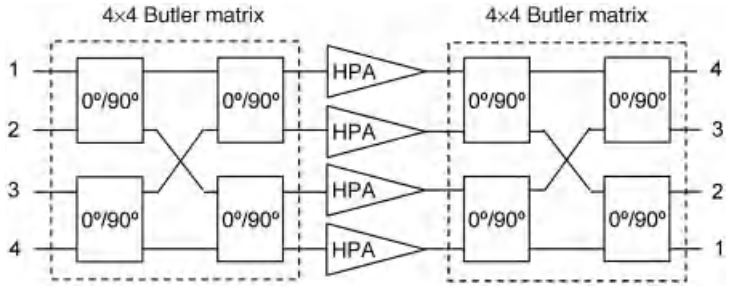


FIGURE 6.8 Four-by-four multiport amplifier with input and output ports labeled (after Mallet et al. (2006)). Redundancy not shown.

in gain and phase and must maintain the matching, that is, **track**, over temperature, power, and life or the scheme with the Butler matrices will not work well; there will be residual amounts of other signals in the outputs. This architecture is also used for SSPAs (Section 6.5.1.1), hence the figure label "HPA" instead of "TWTA."

Figure 6.9 shows two payload architectures that use TWTA-based MPAs. The first is for a reflector-based antenna with a feed cluster (Egami and Kawai, 1987). Each MPA amplifies signals for a few feeds. The second is for an antenna array, where each MPA amplifies signals for a few radiating elements which together form one beam (Mallet et al., 2006).

6.4.2.2 Combining Sometimes **TWTA-combining** is implemented to create a greater P_{out} than one TWTA alone can produce. The RF signal goes through one (L)CAMP and is then divided to go to the TWTAs, which are matched as much as possible in their nonlinearity performance. Then the signals are combined. Two common combining architectures are shown in Figure 6.10. The Sirius Radio satellites incorporate large-scale combining, since each satellite combines all 32 TWTAs in 16 dual TWTAs (Briskman and Prevaux, 2001).

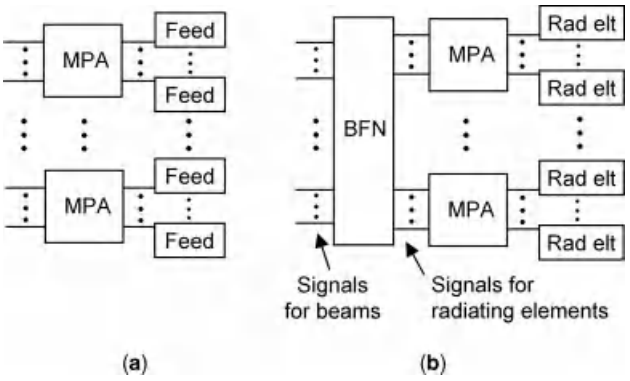


FIGURE 6.9 Payload architectures that use TWTA-based MPAs.

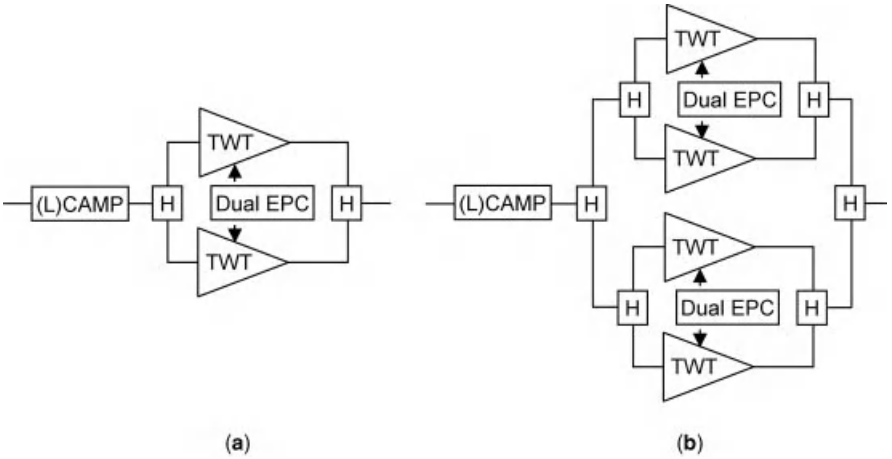


FIGURE 6.10 Examples of TWTA combining (a) dual TWTA combined and (b) two dual TWTAs combined.

Not just the TWTAs must track over temperature but also the output waveguide or coax. The lower the frequency, the easier it is to combine TWTAs, because the longer guided wavelength makes it easier to match waveguide and coax lengths. See Section 8.4.3 for a tip on how to ease combining and Sections 4.3.1 and 4.3.2 respectively for coax and waveguide behavior over temperature.

6.4.2.3 Redundancy Scheme In traditional payload architecture, typically all the TWTAs of one frequency band, for example, Ku-band, belong to one redundancy ring (Section 4.6.2), in which they back each other up in case of failures. The primary, that is, first-choice, TWTA for any particular channel in the band has been optimized for that channel. The first-redundant TWTA for the channel, that is, the first choice to replace the primary TWTA if it fails, may also have been optimized for the same channel. But for a payload that carries a lot of channels, this is not often the case for second- and higher order-redundant TWTAs, which have probably been optimized for a larger bandwidth that covers all the channels it is likely to back up. If the second-redundant TWTA has to be employed at some point in payload life, its performance will not be quite as good as the primary TWTA's was. Sometimes the payload specifications for a channel are required to hold when the primary TWTA is used and when the first-redundant TWTA is used.

The question remains, what the redundancy scheme for the (L)CAMPs is. In the case where one (L)CAMP serves one or two TWTAs, the (L)CAMP is in the redundancy ring along with its corresponding TWTA(s). In the case where the (L)CAMP feeds many TWTAs, it is in its own separate redundancy ring, as in the Sirius Radio satellites, where the LCAMP has 3:1 redundancy (Briskman and Prevaux, 2001).

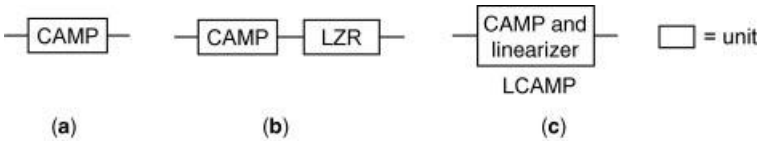


FIGURE 6.11 (L)CAMP realizations as units.

6.4.3 TWTA Subsystem Architecture

The TWTA subsystem has a rather wide range of embodiments and terminology to describe them, besides the typical ones given in Section 6.4.1. The particular embodiments selected for a payload depend on spacecraft-level considerations.

The (L)CAMP possibilities are illustrated in Figure 6.11.

The TWT and EPC are sold in various unit combinations, illustrated in Figure 6.12. There are the single and dual TWTAs and the single and dual **microwave power modules (MPMs)**. The MPM includes the LCAMP function (Tesat, 2007). The **dual TWTA** includes a **dual EPC** to power the two TWTs. The dual EPC is not two EPCs but has only 1.1 times the parts count of a single EPC, at least from one manufacturer (Phelps, 2008). There is a rare EPC failure that brings down both TWTs, but most of the EPC failures bring down only one TWT. The reliability of each TWTA in the dual TWTA is only a little lower than the reliability of a single TWTA, while the mass of the former is less than that of the latter (Phelps, 2008). The two TWTs that go into the dual TWTA are chosen as a matching set in their nonlinear performance.

Figure 6.13 shows ways in which the units of a TWTA subsystem can be integrated. The examples in Figure 6.13a–c each represent one subsystem, while the example in Figure 6.13d represents two.

6.4.4 Channel Amplifier

6.4.4.1 (L)CAMP Unit Architecture and Technology The architecture of a typical LCAMP unit is shown in Figure 6.14 (after Khilla et al. (2002)). A CAMP, missing the linearizer, would also be missing the linearizer-associated circuitry. This particular LCAMP is not part of a flexible TWTA subsystem (Section 6.4.8).

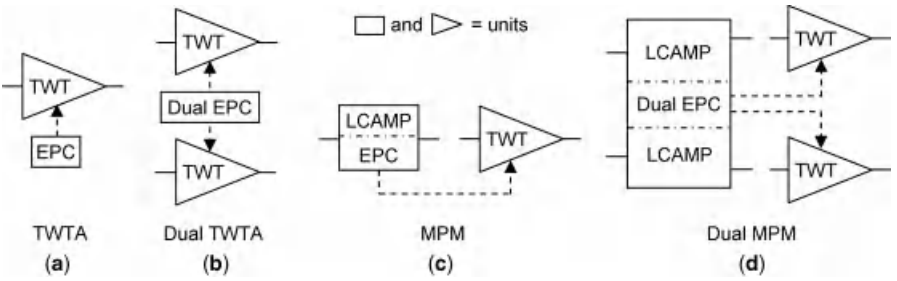


FIGURE 6.12 TWTA and microwave power module realizations as units.

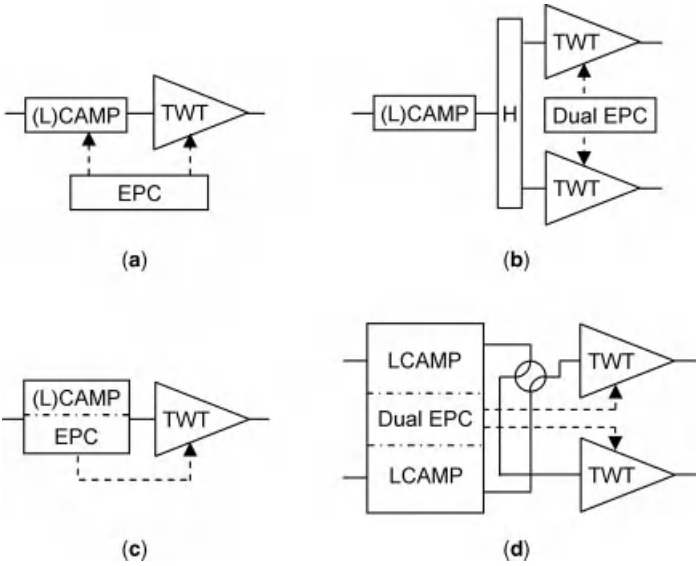


FIGURE 6.13 Examples of TWTA subsystem architecture (a–c) single subsystems, (d) two subsystems.

The unit has two modules, the RF module and the control-and-DC module. The RF module has three variable-gain amplifier (VGA) sections. Immediately after the first VGA section is the first power detector, whose reading the control module uses in ALC mode (this feature does not come from Khilla et al. (2002) but from Thales (2005)). The control module commands both the first and second VGA sections differently, depending on whether the LCAMP is in ALC mode or FGM (also Section 6.4.1). After the first detector but still before the second VGA section is the linearizer. Between the second and third VGA sections is a second power detector, whose reading the control module uses to prevent overdriving the TWTA; if the

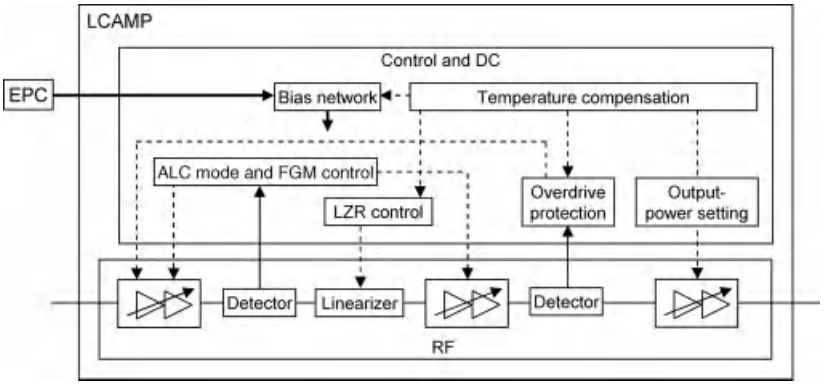


FIGURE 6.14 LCAMP unit architecture example (mostly after Khilla et al. (2002)).

TABLE 6.2 Example of Parameters in a CAMP Specification

Mode	Parameter in CAMP Specification	Units
Both FGM and ALC mode	Operational frequency range	GHz
	RF input-drive level to saturate TWTA, min and max	dBm
	Noise figure—at max gain, min gain	dB
	Return loss—input, output	dB
	Operational temperature range	°C
FGM	Commandable gain range and step size	dB
	Gain variation—over 36 MHz, over full band	dB pk-pk
	Gain stability at any frequency—over 15°, over operational temperature range; from aging and radiation	dB pk-pk
ALC mode	Commandable output-level range and step size	dB
	Output power variation—over 36 MHz, over full band	dB pk-pk
	Output power stability at any frequency—over 15°; from aging and radiation	dB pk-pk
	ALC time constant (Section 10.5.2)	ms

reading is too high the control turns down the gain in the first VGA section. The last element in the RF module is the third VGA section. The control-and-DC module creates the biases for the active elements. It temperature-compensates the bias network, the linearizer, the overdrive protection, and the output power.

The LCAMP technology is microwave hybrid ICs packaging (Khilla and Leucht, 2001; Thales, 2006). The RF module contains MMICs (Khilla and Leucht, 2001; Thales, 2006; Comparini et al., 1999).

6.4.4.2 CAMP Specification Table 6.2 gives an example of the parameters in a CAMP specification. For a LCAMP, the nonlinear performance of the linearizer is not specified with the CAMP but together with the TWTA.

6.4.5 Linearizer

6.4.5.1 Linearizer Architecture and Technology If there is a predistortion linearizer in the TWTA subsystem, it may have an architecture similar to that shown in Figure 6.15a. The linearizer has a "bridge" structure, in that it has linear and nonlinear arms combined at input and output with hybrid couplers (Khilla, 2011b). The linear arm contains a phase shifter and a delay line to equalize the delay in the two arms. The nonlinear arm contains a distortion generator and an attenuator. The distortion generator generates gain and phase predistortion versus input RF-drive level. The attenuator equalizes the signal level in the two arms. Such a linearizer at L-band is described in Khilla et al. (2002), where the distortion generator is a MESFET. This linearizer has NPR better than 14.5 dB even in overdrive (15 dB being a typical requirement for a backed-off LTWTA). The linearizer must be tuned during manufacture to the particular TWTA(s) it will be used with.

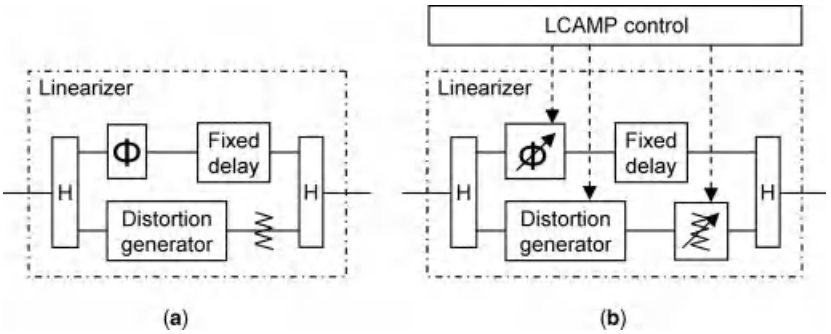


FIGURE 6.15 Linearizer architecture examples with (a) fixed tuning and (b) commandable tuning (latter drawing after Zhang and Yuen (1998)).

If the linearizer may be used for at least two different channels that are separated by a very wide band then the architecture may be similar but with commandable control by the LCAMP's control module, as shown in Figure 6.15b (after Zhang and Yuen (1998)). The linearizer's gain-expansion and phase-advance curves can be adjusted on orbit by modification of the phase shifter and attenuator biases, and the functions' curvatures can be adjusted by modification of the distortion-generator bias (Zhang and Yuen, 1998). This particular linearizer, whose distortion generator uses Schottky diodes, can be used for channels within a 30% bandwidth (Yuen et al., 1999). A similar linearizer is described in Villemazet et al. (2010). An improved, wideband linearizer has a gain-and-phase equalizer in each arm, which allows the linearizer to have varying nonlinear behavior across frequency, which is especially important at double-digit GHz frequencies (Khilia et al., 2011).

Figure 6.16 shows the two different ways that a gain linearizer can work. Figure 6.16a shows a linearizer that reduces the input power to the TWTA by about 2 dB for most of the P_{in} range except near saturation. Figure 6.16b shows a

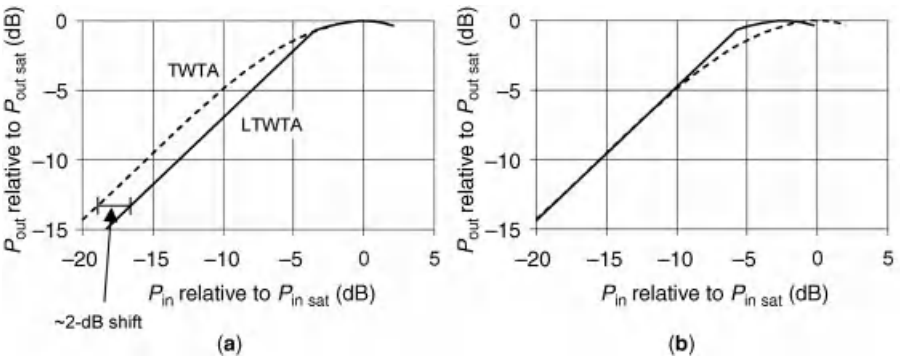


FIGURE 6.16 Two ways that gain linearizer can work.

linearizer that does not affect the input power to the TWTA for most of the P_{in} range but near saturation increases the input power, thus producing **gain expansion**. In both cases, the linearizer increases the P_{in} range over which the LTWTA has nearly linear gain.

On-orbit linearizer technology is a mixture. One L-band linearizer module is a microwave integrated circuit (Khilla et al., 2002). A K-band linearizer module is a microwave hybrid integrated circuit incorporating three MMICs (Khilla and Leucht, 2001). A more recently reported Ku-band linearizer module is a multichip module (see Wikipedia) (Villemazet et al., 2010).

6.4.5.2 LTWTA Nonlinear Performance An example of a TWTA's nonlinear performance was given in Figure 6.1. Here we show, in Figure 6.17, the P_{out} versus P_{in} and phase shift versus P_{in} curves for a linearized TWTA, where the TWTA's behavior before linearization is very much like the TWTA's of Figure 6.1. The LTWTA is close to being a clipper, which is a theoretical device that is linear up to a certain RF input drive and then saturated beyond. This particular linearizer has gain expansion to precompensate the TWTA's gain compression. A small area of minor over-compensation can be seen in Figure 6.17a, at between 8 and 5 dB of input backoff. This is typical. In the region of at least 5-dB input backoff, the linearizer effectively adds about 2 dB to the input P_{in}, as illustrated in Figure 6.16. The phase shift from the LTWTA is almost an order of magnitude smaller than that of the TWTA; in this example, the maximum phase shift difference between the small-signal regime and 2-dB overdrive is only about 7.5°. State-of-the-art LTWTAs produced by one leading manufacturer have a phase shift at saturation within ±9° of the small-signal phase shift (Khilla, 2011a).

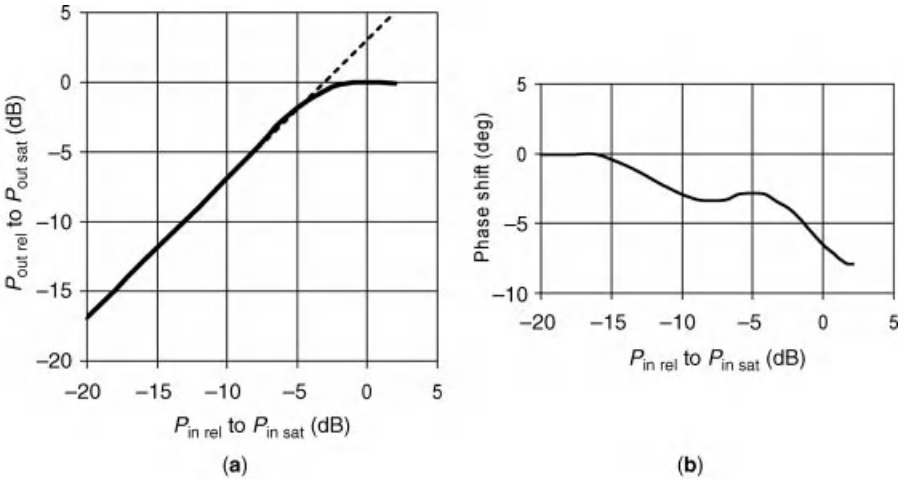


FIGURE 6.17 Example of LTWTA's (a) P_{out} versus P_{in} and (b) phase shift versus P_{in}. (Credit: Tesat-Spacecom.)

6.4.6 TWTA

The TWTA consists of the TWT and the EPC. We describe the EPC briefly first, then the TWT in some detail, and last we give a list of typical parameters in a (L)TWTA specification.

6.4.6.1 Electronic Power Conditioner The EPC performs many functions for the TWT and (L)CAMP. The EPC and the TWT are designed together for best performance (L-3, 2009b). Among other things, the EPC provides DC voltages to the cathode heater, cathode, anode, focusing electrode, helix, and the collector stages (see Section 4.6.2). Common S-band and C-band cathode voltages are 4 kV, Ku-band are 6 kV, and Ka-band and above are 8–14 kV (Barker et al., 2005). The EPC also provides DC power to the (L)CAMP.

The two leading global TWTA suppliers have reported their EPC power efficiency as 94% (L-3, 2009a) and 95–96% (Braetz, 2011b), respectively.

6.4.6.2 TWT Architecture and Technology The TWT is a complex and touchy device to design, is extremely demanding to build, and takes many months to manufacture. Typically the procurement of the TWTAs drives the payload schedule, so the satellite manufacturer places the order at the very beginning of the satellite program, even before their output powers are known exactly. A TWTA is the most expensive unit of a nonprocessing payload. For these reasons, some payload engineers and spacecraft customers become involved in discussions with the TWT manufacturer, for which it is very helpful to have a broad understanding of the TWT, which we aim to give here.

An excellent exposition on the TWT is given in Barker et al. (2005). A shorter but confusingly organized exposition is L-3 (2009b). The latter does contain, however, a description of the full set of TWTA performance parameters.

Figure 6.18 is a drawing of the vacuum assembly of a typical helix TWT (Feicht et al., 2007). There are three connected cylindrical sections with the thinnest one in

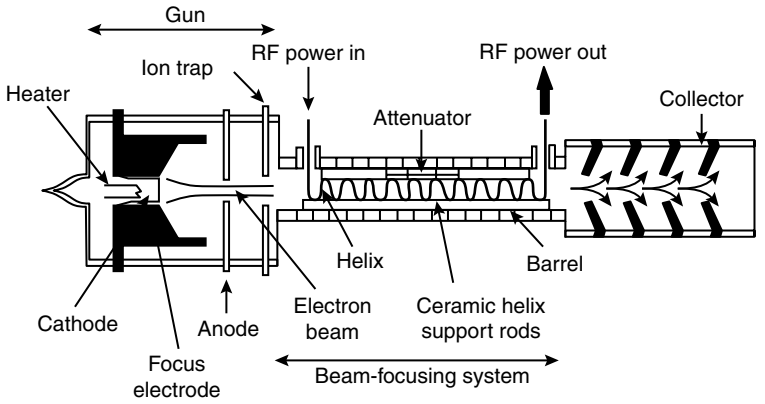


FIGURE 6.18 Drawing of typical helix-TWT vacuum assembly. (Image courtesy of US Air Force, from Feicht et al. (2007)).

the middle, the tube containing the helix. In brief, how the TWTA works is that the electron gun of the first cylindrical section shoots a high-power beam of electrons down the tube, through the center of the helix. The RF signal is launched into the helix. The signal travels at the speed of light down the helix wire but, because of the helix windings, its speed down the tube's axis is slow enough to match that of the electron beam. The RF electric field and the beam interact, causing the RF signal to pick up power from the beam. The amplified RF signal is coupled out of the helix. The collector of the third cylindrical section collects the "spent" electrons (Thales, 2001).

The TWT vacuum assembly is enclosed in a package that serves several purposes: mechanical support, thermal path (including the baseplate) for conduction of waste heat, electromagnetic-interference shield, and protection for the magnets and high-voltage connections (L-3, 2009b).

We now describe the TWT in more detail, first the electron-beam aspects (gun, tube, and collector) then the RF aspects, with thermal considerations of both. Last is a description of how the RF signal and electron beam interact to produce the amplification.

The gun has a cathode, which ejects electrons. The anode has a higher voltage potential than the cathode so it strongly attracts the electrons, which speed up and pass through the middle of the anode. The focus electrode aids in proper formation of the electron beam (L-3, 2009b). In even more detail, the cathode comes in two types, the M-type and the MM-type (Thales, 2001). The M-type is the more common (Barker et al., 2005). Both cathodes are made of a porous tungsten matrix (making them "dispenser cathodes" (L-3, 2009b)) impregnated with barium oxide (Thales, 2001) or a compound of other metal (L-3, 2009b). The barium migrates to the emitting surface of the cathode, and the cathode's high temperature causes the barium to emit electrons (Thales, 2001). The M-type (metal-coated) cathode is coated with osmium, while the MM-type (mixed-metal) is coated with tungsten and osmium. The coating lowers the temperature required for the cathode to emit electrons to "only" about 1000°C (Thales, 2001). The gun sets the TWTA's noise figure (Limburg, 1997), which is typically 25–35 dB for 50–150 W TWTAs (Barker et al., 2005).

The electrons are held in a tight beam down the length of the tube, in order that they go through the middle of the helix, ideally without touching it. This beam "focusing" is accomplished by a series of annular magnets that runs the length of the tube. The magnets are made of samarium cobalt almost exclusively (L-3, 2009b). They are assembled with alternating magnetic orientation, as shown in Figure 6.19a, producing "periodic permanent magnetic (PPM) focusing." The structure that holds the magnets is made of iron "pole pieces" and nonmagnetic spacers brazed together (Karsten and Wertman, 1994; Thales, 2001). The inner surface of this structure is the "barrel" and forms part of the "vacuum envelope" of the vacuum assembly. The helix does not touch the barrel. Figure 6.19b shows a detail of an iron pole piece and a magnet. The outside of the PPM structure has cooling fins (not shown) to conduct heat to the baseplate.

The spent electrons, which still have most of their energy, must be collected. The space TWT's collector has typically four stages (L-3, 2009b), where the first stage slows down and collects the lowest-energy electrons, the second stage the next

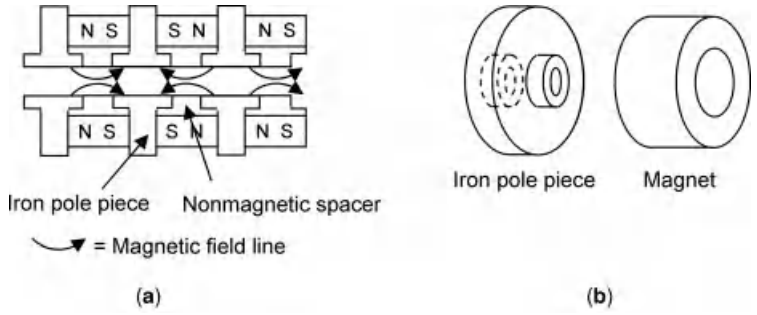


FIGURE 6.19 TWT's periodic-permanent-magnet stack: (a) part of stack showing alternating magnet orientation and magnetic field lines and (b) pole piece and magnet.

lowest-energy electrons, and so on. The electrons are slowed down by the negative voltage of the collector compared to the tube, the latter being at ground potential; such a collector is "depressed." The second collector stage has a lower voltage than the first stage does, the third lower than the second stage, and so on. The repelling force among the electrons spreads them out before they hit the collector walls (L-3, 2009b). The greatest part of the heat in the TWT to be dissipated, is in the collector. There are two ways to dissipate the heat. One is to thermally couple the collector to a baseplate, which is in contact with a spacecraft heat pipe, which conducts away the heat. The other way is to couple the collector to large attached fins or stacked cups, which radiate the heat into space (Thales, 2001). In the latter case the TWTs are located as usual on the inside surface of a spacecraft panel but the radiators protrude beyond the panel into space.

The heat from the TWT is dissipated all or partly via conduction from the TWT baseplate to the spacecraft heat pipes (Section 2.2.1.6). When the heat is dissipated entirely by conduction, the TWT is "conduction-cooled (CC)." When the heat is partly dissipated by radiating cups or fins attached to the TWT collector, which add mass to the TWT, the TWT is "direct-radiation-cooled (DRC)." An example of a 110 W C-band TWTA shows that if the TWT is DRC, only half as much heat has to be conducted away as when it is CC (Barker et al., 2005).

Now for the RF part of the TWT. The RF is coupled into the helix wire. The helix windings slow down the axial (i.e., down the length of the tube) phase velocity of the RF wave sufficiently to establish synchronism with the electron beam, making the helix a "slow-wave structure." The helix slow-wave structure is inherently wideband (L-3, 2009b). At the end of the helix, the amplified RF is coupled off. The helix is at ground potential. It is supported by three ceramic rods that run the length of the tube. The rods hold the helix in the middle of the barrel (Thales, 2001). They also serve to dissipate to the barrel the heat generated in the helix from RF losses and electron impacts (L-3, 2009b). An alternative but much less common slow-wave structure is the coupled cavity, described in L-3 (2009b). The coupled-cavity TWT has a narrower bandwidth but higher power than the helix TWT does. It will not be further discussed.

If there are at least two impedance mismatches along the RF path, for example, at the RF input and output couplers, an oscillation can arise. The most common means to prevent this is to introduce a "sever," that is, "cutter," about halfway down the length of the helix (L-3, 2009b). Most TWTs have one sever, but some have two or none. The sever attenuates the RF signal entirely; the information of the RF signal is carried across the cut by the electron beam. The sever itself is also a potential impedance discontinuity. In some TWTs the sever is distributed, consisting of carbon deposited on the ceramic rods, with the amount of carbon being the largest about in the middle of the tube (L-3, 2009b). TWTs at 30 GHz and above can experience "backward-wave oscillations" (Barker et al., 2005), which occur out-of-band but rob power from the intended output signal. A distributed sever is more effective than a discrete one for suppressing this (Nusinovich et al., 1998).

The electron beam interacts with the axial component of the RF electric field (Barker et al., 2005). In the first region of interaction, the electrons "bunch." When individual electrons enter the electric field, depending on the phase of the field polarity, they are either sped up or slowed down. They thus form bunches. In the second region of interaction, the bunches contribute power to the RF signal. The bunches are in decelerating regions of the electric field, so the bunches are continually slowed down. They transfer their kinetic energy to the RF wave by inducing a modified wave onto the original wave. The RF signal gets amplified exponentially. At some point the bunches start to come apart or they can no longer be kept in synchronism with the RF wave. This is where the helix ends. The area of synchronism is often extended by "tapering" the end of the helix, that is, decreasing the helix pitch to further slow down the RF wave. When the TWT is operated at saturation, the electron bunches at helix output have a 30–50° phase delay relative to when the TWT is operated at small-signal. The bunches "pull" the RF wave back by this same amount (Barker et al., 2005).

6.4.6.3 (L)TWTA Specification Table 6.3 gives an example of the specification parameters for a (L)TWTA. This particular (L)TWTA will be used over a range of IBOs, so its nonlinearity is specified by masks over IBO. A (L)TWTA to be used only for multicarrier operation could instead be specified on NPR (see Section 6.2.2 for discussion on nonlinearity specification). One parameter we have not seen before is the out-of-band rejection at the uplink frequencies, to minimize any power that could potentially sneak back into the receiver.

6.4.7 TWTA Subsystem Performance

Table 6.4 summarizes the current performance of TWTA subsystems for commercial payloads, made by of one of the leading global subsystem suppliers. The performance is given in terms of the key parameters output power, bandwidth, and power efficiency. The data is all at CW saturation. The other leading supplier's performance is similar (L-3, 2009a).

TABLE 6.3 Example of Parameters in a (L)TWTA Specification

Parameter in (L)TWTA Specification	Units
Operational frequency range	GHz
CW-saturation output power, minimum across frequency range at BOL	W
CW-saturation output-power stability—over 15°, over full operational temperature range; over life	dB
RF input-drive level for CW to saturate—BOL, EOL	dBm
OBO versus IBO curve	Mask
Phase shift versus IBO curve	Mask
Gain flatness—on 36 MHz, on full band	dB pk-pk
Gain slope—at CW sat, at small-signal	dB/MHz
Phase deviation from linear versus frequency, at small-signal	° pk-pk
Second- and third-harmonic output power at CW saturation	
Spurious outputs—in-band and noncoherent out-of-band; saturating CW and no RF drive	dBc
Out-of-band rejection relative to in-band gain, at uplink frequencies	dB
Noise figure	dB
PAE at operating point	%
Return loss—input, output	dB
Operational temperature range	°C

6.4.8 Flexible TWTA Subsystem

A flexible payload-level TWTA architecture has been described in Section 6.4.2.1. Here we describe a flexible TWTA subsystem which can alter $P_{\text{out sat}}$, so that as on orbit the required P_{out} changes, the TWTA can be kept operating at its most power-efficient (Khilla et al., 2005; Khilla, 2008). The approach hinges on the fact that

TABLE 6.4 Summary of One Supplier's TWTA Subsystem Performance[a] for CW

	High-Power			Low-Power		
Frequency Band	P_{sat} (W)	Bandwidth (MHz)	Typical Power Efficiency (%)	P_{sat} (W)	Bandwidth (MHz)	Typical Power Efficiency (%)
L-band	250	100	59	70	50	56
S-band	275	100	62	70	100	60
C-band	125	350	64–66	20	350	55–57
X-band	160	500	62	12	500	51
Ku-band	220	1000	64	25	2050	59
K-band	130	2000	61	15	2000	46–53
Ka-band	60	500	47	10	500	46

[a]TWT data from Thales (2009); assumed 95% EPC PAE from Braetz (2011b); assumed 2 W DC power consumption for (L)CAMP.

changing a TWT's anode voltage or, equivalently, its cathode current causes $P_{\text{out sat}}$ to change. As $P_{\text{out sat}}$ is reduced by 1 dB, the TWT's gain reduces by about 5 dB and gain slope versus frequency arises. The LCAMP commands the EPC to change the cathode current and at the same time puts out a different power level to the TWT and compensates the gain slope. The LCAMP has an additional, medium-power amplifier in its output amplification section. $P_{\text{out sat}}$ can be set to within 0.1 dB over a range of 4 dB. For a 4-dB reduction in $P_{\text{out sat}}$, the DC power consumption is reduced by 24% and heat dissipation by 30% relative to no $P_{\text{out sat}}$ reduction. Such subsystems were slated to be launched in 2009 (Jaumann, 2009).

6.4.9 TWTA Subsystem Environmental

The (L)CAMP, the TWT, and the EPC change over life, and the (L)CAMP is designed to compensate the changes in all three (Khilla et al., 2002). The categories of environmental influence are temperature, radiation, and aging. Aging and radiation are sometimes combined to form the category "life."

6.4.9.1 Temperature The (L)CAMP would have decreasing gain with higher temperature, without temperature compensation. However, the unit has a temperature sensor, whose readings the control module uses to adjust both the input and output amplifier sections (see Figure 6.14) (Khilla et al., 2002).

For a constant RF drive level into the (L)CAMP, as the TWTA temperature rises the (L)CAMP increases its output power to the TWTA to compensate the TWTA (Khilla et al., 2002).

6.4.9.2 Radiation The (L)CAMP's gain decreases over life, due to its component amplifiers' exposure to radiation. This is compensated by adjustment of the output amplifier section (Khilla et al., 2002).

The TWT is intrinsically radiation-hard, while the solid-state power-conditioning circuits in the EPC require radiation protection (Barker et al., 2005).

6.4.9.3 Aging The (L)CAMP's gain decreases over life, due to its component amplifiers aging. It compensates itself by adjusting its output amplifier section (Khilla et al., 2002).

One of the two leading global TWTA suppliers has reported a phenomenon of its TWTs occurring since 1977, that its TWTs' gain increases through end of spacecraft life (Feicht et al., 2007). This applies to the saturated gain as well as the small-signal gain, namely, the P_{out} versus P_{in} curve shifts upward over life (Goebel et al., 2003). The supplier has observed this phenomenon on its TWTs from S-band to Ka-band. The cause is the degradation of the sever material on the ceramic support rods. The rate of the gain increase is largest right after the TWT is fabricated, decreases exponentially during a few thousand hours of operation, then becomes linear. The linear gain growth starts some time after the end of burn-in, a part of the fabrication process meant to stabilize the TWT. The linear gain growth therefore applies to on-orbit operation and can be a problem in interference scenarios. The

slope of the linear gain increase is larger and more variable, the higher the TWT's saturated output power. The supplier reduced the size of the problem starting in 1998 (Goebel et al., 2003). The (L)CAMP can attempt to compensate this gain increase by adjustment of the output amplifier section.

The other leading TWTA supplier says it has no such reports or customer feedback and no such results of ground testing on the TWTs that it procures (Braetz, 2011a). The TWT aging is believed to be fully compensated by the EPC (Braetz, 2011c).

6.5 SOLID-STATE POWER AMPLIFIER

When the frequency band is L-, S-, or C-band and the transmit antenna is a phased array or a reflector-based antenna with a phased-array feed, SSPAs are used. The deciding factor here is the small size of the SSPA compared to the TWTA. The SSPA is not ordinarily used for other payload applications because of its low power efficiency compared to the TWTA. See the comparison in Section 6.3 above between the SSPA and the TWTA.

The SSPA is different from the TWTA subsystem in that the SSPA unit contains the preamplifier, linearizer if any, and EPC. (An exception is the Globalstar-1 SSPAs that feed a multibeam phased array for the user downlinks: the EPC is separate and serves all 91 SSPAs (Metzen, 2000).)

6.5.1 SSPAs in Payload

6.5.1.1 Architecture in Payload, Traditional and Flexible Figure 6.20 shows the transmitter side of three payload architectures that incorporate SSPAs. All have antenna arrays (Section 3.5).

a. In a traditional architecture, the antenna array's beam-forming network (BFN) (Section 3.5.6) outputs one signal per radiating element. A SSPA powers each signal and is built in behind the radiating element. This arrangement of SSPAs and radiating elements is an active antenna array. Examples of this architecture are the Globalstar-1 S-band transmitter for user links (Hirshfield, 1995) and the Iridium L-band transmitter for user links (Schuss et al., 1999).
b. In one flexible architecture, the BFN outputs one signal per radiating element. The signals are grouped, and an MPA powers each group (Section 6.4.2.1). Examples of this architecture are the Globalstar-2 S-band transmitter for user links (Darbandi et al., 2008), the Inmarsat-4 L-band transmitter for user links series of satellites (Seymour, 2000), and the N-Star S-band transmitter for user links (Tanaka et al., 1996). N-Star A is the first in a series of Japanese satellites for domestic mobile services, launched from 1995 to 2002 (Gunter's, 2011).
c. Another flexible architecture is similar to (b) except that the BFN incorporates the input Butler matrices of the MPAs (Roederer, 2010). The collection of

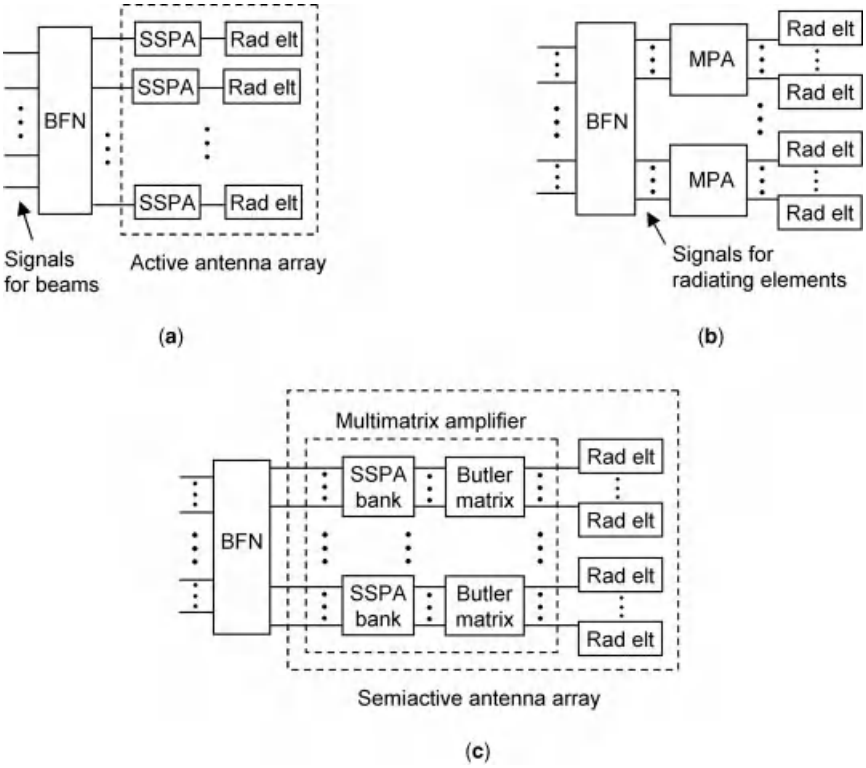


FIGURE 6.20 SSPA architectures in payload: SSPAs (a) directly feeding active antenna array, (b) arranged in MPAs feeding antenna array, and (c) arranged in multimatrix amplifier feeding semiactive antenna array.

SSPA banks and output Butler matrices is a **multimatrix amplifier**. The multimatrix amplifier together with the radiating elements is a semiactive antenna array (Section 3.5.7). Examples of this architecture are the L-band transmitter for the user links of Inmarsat-3 satellites (Perrott and Griffin, 1991) and the L-band transmitter for user links of Thuraya satellites (Roederer, 2010). Thuraya satellites were launched from 2003 to 2008.

The advantage of the MPA and multimatrix architectures is that the combined power of the SSPAs in a bank is available all to just one signal, equally to all, or anything in between, according to the relative proportions of input signal strengths. So these architectures are flexible in that they can accommodate fluctuating and unbalanced beam traffic. Potential disadvantages are that the signals must all be at different frequencies and that the SSPAs must always be backed off to the multicarrier region (Mallet et al., 2006).

In all of these architectures, it is imperative that the SSPAs exhibit gain and phase matching. Since this matching is always needed over temperature and for the

MPA and multimatrix amplifier also over input RF drive, the SSPAs must also exhibit gain and phase tracking over the changing conditions.

6.5.1.2 Redundancy Scheme The possibility of SSPA failure over the payload lifetime must be accounted for. In the case of the active antenna array, the array is designed so that the loss of a few active radiating elements does not significantly degrade the beam patterns. In the cases of the SSPAs being arranged in MPAs or multimatrix amplifiers, the SSPAs are not necessarily redundant, either. The failure of one SSPA in an 8×8 MPA causes a signal loss of 1.2 dB and degrades the isolation to 18 dB. The numbers are more benign for a larger MPA and less benign for a smaller one (Egami and Kawai, 1987). If such numbers are acceptable, the SSPAs need not be redundant.

The Eurostar generic architecture for geomobile satellite payloads employs redundancy rings and has as well a soft response to additional failures. When the payload loses an entire active feed chain, the ground computes new beam weights and uploads them, and the onboard digital processor implements them. This architecture has been applied to Inmarsat-4 (Mallison and Robson, 2001).

6.5.2 SSPA Unit Architecture and Technology

Figure 6.21 gives an example of an SSPA unit architecture, from Inmarsat-4. This SSPA adjusts itself based on the RF input power it measures. It has no ALC, although an ALC is common among SSPAs. Like most other SSPAs, this SSPA has low-power, medium-power, and high-power amplification sections. The high-power section usually consists of amplifiers in parallel whose outputs are combined. The low-power section corresponds most closely to a (L)CAMP, and the medium-power and high-power amplification sections together correspond most closely to a TWTA, in terms of function and the amount of gain each has.

Currently two technologies are used in the high-power amplifier section, GaAs pHEMT and GaN HEMT (Katz and Franco, 2010). Both are FETs. These technologies are described in more detail in Section 5.4.1. The standard technology is the

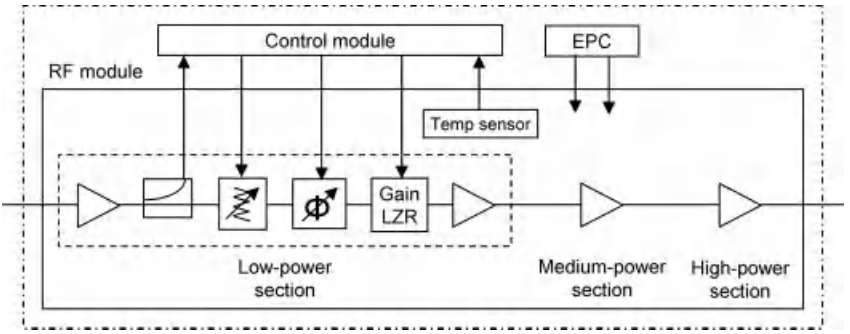


FIGURE 6.21 SSPA unit architecture, Inmarsat-4 L-band example (after Seymour (2000)).

GaAs pHEMT, which is built on or grown on top of a GaAs substrate. Such devices are used in the following SSPAs:

- Globalstar-1 for S-band user links (Hirshfield, 1995) and C-band gateway links (Ono et al., 1996)
- Iridium for L-band user links (Schuss et al., 1999) and K-band crosslinks and gateway links (Agar et al., 1995)
- Inmarsat-4 for L-band user links (Seymour, 2000) and C-band gateway links (Kiyohara et al., 2003).

The new technology is the GaN HEMT, which has a SiC (silicon carbide) substrate and a GaN channel. The devices have a much higher bandwidth than GaAs devices (Katz and Franco, 2010). SSPAs with such devices should be possible for frequencies up to W-band (Barnes et al., 2005).

6.5.3 Linearized SSPA

In some payload applications the nonlinear performance of a SSPA without linearizer is sufficient. A performance example was shown in Figure 6.2, for the SSPA that is part of the Inmarsat-3 L/C-band upconverter/SSPA unit. At the 1.6-dB compression point, the phase shift is 9° (Khilla and Leucht, 1996). In general, the phase shift of an unlinearized SSPA at the 2-dB compression point P2dB and the phase shift of a LTWTA at saturation have about the same magnitude (Khilla, 2011a).

In other applications linearization is needed to meet high requirements, and the high linearity allows the SSPA to operate with higher efficiency. Linearization is implemented in the low-power amplification part of the SSPA. Examples are the Globalstar-1 SSPAs for the C-band gateway links (Ono et al., 1996), the Inmarsat-3 SSPAs for the C-band fixed-terminal links (Khilla and Leucht, 1996), and the Inmarsat-4 SSPAs for the C-band gateway links (AIAA JSFC, 2003). All these SSPAs amplify multiple carriers.

The gain linearization in a SSPA works in either of the same two ways as for a LTWTA (Section 6.4.5.2). The method shown in Figure 6.16b is used in the Inmarsat-3 C-band SSPA.

Figure 6.22 shows the linearized performance of the same SSPA. Over the range of 20 to 0 dB IBO, where 0-dB IBO is the P_{in} at which unlinearized gain compression is 1.6 dB, the linearized phase shift is within $\pm 1.5°$ and the gain compression is within ± 0.25 dB. The multicarrier NPR specification was 23 dB; linearization allows the SSPA to be operated 2 dB closer to the 1.6-dB compression point (Khilla and Leucht, 1996).

6.5.4 Flexible SSPA

Besides the flexible ways in which the payload architecture can use SSPAs (Section 6.5.1.1), the SSPA unit itself can also be designed to be flexible. A payload that implements both kinds of flexibility is Globalstar-2, in its S-band architecture with MPA (Darbandi et al., 2008).

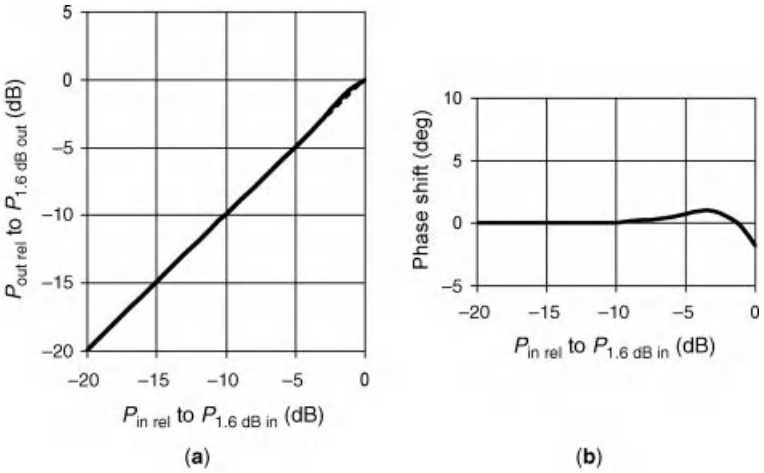


FIGURE 6.22 Example of linearized SSPA's (a) P_{out} versus P_{in} curve and (b) phase shift versus P_{in} curve (after Khilla and Leucht (1996)).

The SSPA unit itself can have two different kinds of flexibility designed into it. Some SSPAs have both. Here we do not count an ALC capability as a flexibility element, because it is so common among SSPAs. The two kinds of flexibility are as follows:

- Gain and phase compensation over input RF-drive level, accomplished in the low-power-amplification section of the SSPA. Examples of this are Inmarsat-4's L-band user links (Seymour, 2000) and C-band gateway links (Kiyohara et al., 2003).
- Automatic rebiasing of the SSPA in order to change its output power in response to detected input RF-drive level. This maintains good power efficiency even though the traffic level varies. There is also gain and phase compensation to keep the gain and phase constant. Examples of this are Globalstar-1's S-band user links (Metzen, 2000) and C-band gateway links (Ono et al., 1996), as well as Globalstar-2's S-band user links (Darbandi et al., 2008).

6.5.5 SSPA Environmental

The categories of environmental influence are temperature, radiation, and aging.

As the SSPA's temperature increases, its noise figure increases and its gain drops (Microwave Encyclopedia, 2010). To limit this, the SSPA is installed on a heat pipe (Section 2.2.1.6) to carry away the heat. At low temperatures, a heater may be necessary when the SSPA is amplifying to a low power level, for example, the Iridium L-band (Schuss et al., 1999). The SSPA in an antenna array employs temperature compensation of gain and usually phase, to keep all the SSPAs tracking. The BFN needs this, as well as any MPA or MMA.

The GaN HEMT has a much higher operating temperature than the GaAs pHEMT does, so it is expected for GaN that the thermal-control subsystem can be much reduced (Muraro et al., 2010).

GaAs devices must be shielded against radiation, unlike GaN devices, which are radiation-hard (Barnes et al., 2005).

Over life, there are at least two different aging effects. The GaAs device's gain gradually decreases over life even with low RF drive (Chou et al., 2004). For such devices run long-term at saturation, there is also another degradation of gain as well as of P_{out} and power efficiency. For example, a saturating-RF lifetest of a 20 GHz, two-stage MMIC power amplifier resulted in the entire P_{out} versus P_{in} curve shifting down by the same amount, while power efficiency decreased by almost 2 percentage points (Chou et al., 2005). A similar result regarding P_{out} versus P_{in} was observed in pHEMT MMIC power amplifiers designed for the band from 6 to 18 GHz (Anderson et al., 2000).

6.5.6 SSPA Specification

Table 6.5 gives an example of the specification parameters for a SSPA that is not part of an active transmit phased array, that is, where no phase or gain matching is required. It is for multicarrier use since NPR is specified. A typical NPR for

TABLE 6.5 Example of Parameters in a SSPA Specification

Mode	Parameter in SSPA Specification	Units
Both FGM and ALC mode	Operational frequency range	GHz
	NPR	dB
	Noise figure—at max gain, min gain	dB
	Return loss—input, output	dB
	Spurious outputs—in-band and noncoherent out-of-band	dBc
	Out-of-band rejection relative to in-band gain, at uplink frequencies	dB
	Power efficiency for multiple carriers	%
	Operational temperature range	°C
FGM	Commandable gain range and step size	dB
	Gain flatness at small signal—over 36 MHz, over full band	dB pk-pk
	Gain stability at any frequency—over temperature; from aging and radiation	dB pk-pk
ALC mode	Commandable output-level range and step size	dB
	Output power variation at small signal—over 36 MHz, over full band	dB pk-pk
	Output power stability at any frequency—over temperature; from aging and radiation	dB pk-pk
	ALC time constant (Section 10.5.2)	ms

TABLE 6.6 Summary of GaAs SSPA Performance

		For Multiple Carriers		For CW		
Frequency Band	Bandwidth (MHz)	P_{out} (W)	Power Efficiency (%)	P_{out} (W)	Power Efficiency (%)	Reference
L-band	34	15	32	–	–	Astrium, 2011
L-band	34	24	–	40	–	NEC, 2011
L-band	40	–	–	55	–	NEC, 2011
S-band	200	20	35	40	55	Darbandi et al., 2008
C-band	300	–	–	60	47	AIAA JFSC, 2003
C-band	178	25	19	–	–	Ono et al., 1996

multicarrier use is 15 dB. This SSPA has both FGM and ALC modes. A SSPA has a very much lower noise figure than a TWTA does (Barnes et al., 2005). The out-of-band rejection parameter has the purpose of minimizing any power that could potentially sneak back into the receiver.

Table 6.6 summarizes the performance of some current SSPAs in terms of the key parameters bandwidth, output power, and power efficiency. The data for multicarrier use is based on a NPR requirement of about 15 dB. The data for CW use is based on P_{out} of the P2dB point for the S- and C-band SSPAs; this is probably also the case for the L-band SSPAs. The power-transistor technology for all these SSPAs is GaAs. A GaN S-band SSPA has by far the highest CW P_{out}, 150 W, but is not listed in Table 6.6 because it is brand-new (Nakade et al., 2010).

REFERENCES

Agar Jr, BT, Bowser DS, Buer KV, Corman DW, and Grondahl CD (1995). The Iridium K-band MMIC chip set. *Digest, IEEE Microwave and Millimeter-Wave Monolithic Circuits Symposium;* May 15–16; 21–24.

Agilent Technologies (2000). Agilent AN 1287-4, Network analyzer measurements: filter and amplifier examples. Application note. On cp.literature.agilent.com/litweb/pdf/5965-7710.pdf. Accessed Nov. 12, 2010.

AIAA Japan Forum on Satellite Communications (2003). Capital products & review: C-band solid state power amplifier (SSPA). *Space Japan Review*; no. 27; Feb./Mar. On satcom.nict.go.jp/English/english2-3/index.html. Accessed June 27, 2011.

Anderson WT, Roussos JA, and Mittereder JA (2000). Life testing and failure analysis of PHEMT MMICs. *Proceedings of GaAs Reliability Workshop;* Nov. 5; 45–52.

Astrium (2011). SSPA—L band. Product sheet. On www.astrium.eads.net/en/equipment/l-band-sspa.html. Accessed May 25, 2011.

Barker RJ, Booske JH, Luhmann Jr, NC, and Nusinovich GS, editors (2005). *Modern Microwave and Millimeter-Wave Power Electronics*, New Jersey: IEEE Press and John Wiley and Sons.

Barnes AR, Boetti A, Marchand L, and Hopkins J (2005). An overview of microwave component requirements for future space applications. *European Gallium Arsenide and Other Semiconductor Applications Symposium;* Oct. 3–4; 5–12.

Bosch E, Jaeger A, Seppelfeld E, Monsees T, and Nunn RA (2003). TWTA dominance, C-band traveling wave TWTs versus solid state amplifiers. *AIAA International Communications Satellite Systems Conference;* Apr. 15–19; 1–10.

Braetz M, Tesat-Spacecom, Backnang, Germany (2011a). Private communication. Jul. 15.

Braetz M, Tesat-Spacecom, Backnang, Germany (2011b). Private communication. Jul. 19.

Braetz M, Tesat-Spacecom, Backnang, Germany (2011c). Private communication. Nov. 25.

Briskman RD and Prevaux RJ (2001). S-DARS broadcast from inclined, elliptical orbits. *International Astronautical Congress;* Oct. 1–5; 503–518.

Chou YC, Grundbacher R, Leung D, Lai R, Liu PH, Kan Q, Biedenbender M, Wojtowicz M, Eng D, and Oki A (2004). Physical identification of gate metal interdiffusion in GaAs PHEMTs. *IEEE Electron Device Letters;* 25 (Feb.); 64–66.

Chou Y-C, Grundbacher R, Lai R, Allen BR, Osgood B, Sharma A, Kan Q, Leung D, Eng D, Chin P, Block T, and Oki A (2005). Hot carrier effect on power performance in GaAs PHEMT MMIC power amplifiers. *IEEE MTT-S International Microwave Symposium Digest;* June 12–17; 165–168.

Comparini MC, Feudale M, Ranieri P, Suriani A, Gatti G, and Auxemery P (1999). Ka band satellite equipment using European GaAs technology. *Gallium Arsenide Applications Symposium;* Oct. 4–5; pp. 175–180.

Darbandi A, Zoyo M, Touchais JY, and Butel Y (2008). Flexible S-band SSPA for space application. *NASA/ESA Conference on Adaptive Hardware and Systems;* June 22–25; 70–76.

Egami S and Kawai M (1987). An adaptive multiple beam system concept. *IEEE Journal on Selected Areas in Communications;* 5 (May); 630–636.

Feicht JR, Martin RH, and Williams BC (L-3 Communications-Electron Technologies) (2007). (Congressional-Microwave Vacuum Electronics Power Res. Ini.) TWT coatings improvement investigation—TWT gain growth. Final technical report. Feb. 1. Arlington, VA: USAF, AFRL AF Office of Scientific Research. Report no. AFRL-SR-AR-TR-07-0079. Contract no. FA9550-05-C-0173. On www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA463632&Location=U2&doc=GetTRDoc.pdf. Accessed Jul. 6, 2011.

Goebel DM, Schneider AC, Menninger WL, and Weekley JM (2003). Gain increases through end of life in traveling wave tubes. *IEEE Transactions on Electron Devices;* 50; 1117–1124.

Gunter's Space Page (2011). N-Star a, b. March 13. On space.skyrocket.de/doc_sdat/nstar-a.htm. Accessed Jul. 4, 2011.

Hirshfield E (1995). The Globalstar system breakthroughs in efficiency in microwave and signal processing technology. *Proceedings of AIAA/ESA Workshop on International Cooperation in Satellite Communications;* Mar. 27–29; 147–152.

Jaumann G (2009). Improved flexibility by in-orbit-adjustable saturation output power of TWTs. *IEEE International Vacuum Electronics Conference;* Apr. 28–30; 203–204.

Kaliski M (2009). Evaluation of the next steps in satellite high power amplifier technology: flexible TWTAs and GaN SSPAs. *IEEE International Vacuum Electronics Conference;* Apr. 28–30; 211–212.

Karsten KS and Wertman RC, inventors; ITT Corp, assignee (1994). Interlocking periodic permanent magnet assembly for electron tubes and method of making same. U.S. patent 5,334,910. Aug. 2.

Katz A and Franco M (2010). GaN comes of age. *IEEE Microwave Magazine, Supplement;* 11 (Dec.); S24–S34.

Khilla M (2008). In-orbit adjustable microwave power modules MPMs. *ESA Workshop on Advanced Flexible Telecom Payloads;* Nov. 18–20; 1–11.

Khilla A-M, Tesat-Spacecom, Backnang, Germany (2011a). Private communication. Feb. 16.

Khilla A-M, Tesat-Spacecom, Backnang, Germany (2011b). Private communication. Aug. 22.

Khilla A-M, Tesat-Spacecom, Backnang, Germany (2011c). Private communication. Sept. 5.

Khilla A-M, Tesat-Spacecom, Backnang, Germany (2011d). Private communication. Sept. 7.

Khilla M, Gross W, Schreiber H, and Leucht D (2005). Flexible Ka-band LCAMP for in-orbit output power adjustable MPM. *AIAA International Communications Satellite Systems Conference;* Sept. 25–28; 1–12.

Khilla A-M and Leucht D (1996). Linearized L/C-band SSPA/upconverter for mobile communication satellite. *Technical Papers, AIAA International Communications Satellite Systems Conference;* Pt. 1 (Feb. 25–29); 86–93.

Khilla A-M and Leucht D (2001). Advanced Ka-band dual microwave power module, "dual MPM." *Technical Papers, AIAA International Communications Satellite Systems Conference;* Apr. 17–20; 1–10.

Khilia A-M, Leucht D, Gross W, Jutzi W, and Schreiber H, inventors (2011). Predistortion linearizer with bridge topology having an equalizer stage for each bridge arm. U.S. patent application publication 2011/0169566 A1. Jul. 14.

Khilla AM, Scharlewsky D, and Niederbaeumer J (2002). Advanced linearized channel amplifier for L-band-MPM. *AIAA International Communications Satellite Systems Conference;* May 12–15; 1–10.

Kiyohara A, Kazekami Y, Seino K, Tanaka K, Shirasaki K, Fukazawa S, Iwano N, Kittaka Y, and Gill R (2003). Superior tracking performance of C-band solid state power amplifier for Inmarsat-4. *AIAA International Communications Satellite Systems Conference;* Apr. 15–19; 1–9.

Limburg H, Hughes Electron Dynamics (now L-3 Communications Electron Technologies, Inc), Torrance, CA (1997). Private communication. Oct. 21.

L-3 Communications Electron Technologies, Inc (2009a). Space LTWTA products, space qualified TWT product listing; Nov. 18. Space LTWTA products, space qualified EPCs product listing; Nov. 14. On www.l-3com.com/eti/downloads/summarytable_space.pdf. Accessed Jul. 19, 2011.

L-3 Communications Electron Technologies, Inc (2009b). *TWT/TWTA Handbook*, 13th ed. Torrance, CA: L-3 Communications Electron Technologies, Inc.

Mallet A, Anakabe A, Sombrin J, and Rodriguez R (2006). Multiport-amplifier-based architecture versus classical architecture for space telecommunication payloads. *IEEE Transactions on Microwave Theory and Techniques;* 54 (Dec.); 4353–4361.

Mallison MJ and Robson D (2001). Enabling technologies for the Eurostar geomobile satellite. *AIAA International Communications Systems Conference;* Apr. 17–20; 1–10.

Metzen PL (2000). Globalstar satellite phased array antennas. *Proceedings of IEEE International Conference on Phased Array Systems and Technology;* May 21–25; 207–210.

Microwave Encyclopedia (2010). Power amplifiers. Oct. 15. On www.microwaves101.com. Accessed Jul. 1, 2011.

MIL-HDBK-217F (1995). *Military Handbook, Reliability Prediction of Electronic Equipment; Change Notice 2*. Washington: Department of Defense; Feb. 28. On snebulos.mit.edu/projects/reference/MIL-STD/MIL-HDBK-217F-Notice2.pdf. Accessed Jul. 13, 2011.

Muraro J-L, Nicolas G, Nhut DM, Forestier S, Rochette S, Vendier O, Langrez D, Cazaux J-L, and Feudale M (2010). GaN for space application: almost ready for flight. *International Journal of Microwave and Wireless Technologies;* 2; 121–133.

Nakade K, Seino K, Tsuchiko A, and Kanaya J (2010). Development of 150 W S-band GaN solid state power amplifier for satellite use. *Proceedings of Asia-Pacific Microwave Conference;* Dec. 7–10; 127–130.

NEC Toshiba Space Systems Ltd (2011). Solid state power amplifier (SSPA). Product sheets. On www.nec.com/global/solutions/space/satellite_communications/index.html. Accessed Jan. 28, 2011.

Nicol EF, Mangus BJ, and Grebliunas JR (2008). TWTA versus SSPA: analysis update of the Boeing fleet on-orbit reliability data. *IEEE International Vacuum Electronics Conference;* Apr. 22–24; 169–170.

Nusinovich GS, Walter M, and Zhao J (1998). Excitation of backward waves in forward wave amplifiers. *Physical Review E;* 58 (Nov.); 6594–6605.

Ono T, Ozawa T, Kamikokura A, Hayashi R, Seino K, and Hirose H (1996). Linearized C-band SSPA incorporating dynamic bias operation for Globalstar. *Technical Papers, AIAA International Communications Satellite Systems Conference;* Pt. 1 (Feb. 25–29); 123–130.

Perrott RA and Griffin JM (1991). L-band antenna systems design. *IEE Colloquium on Inmarsat-3;* Nov. 21; 5/1–5/7.

Phelps TK (2008). Reliability of dual TWTAs—spacecraft system considerations. *IEEE International Vacuum Electronics Conference;* Apr. 22–24; 173–174.

Roederer AG (2010). Semi-active satellite antenna front-ends: a successful European innovation. *Proceedings of Asia-Pacific Microwave Conference;* Dec. 7–10; 243–246.

Schuss JJ, Upton J, Myers B, Sikina T, Rohwer A, Makridakas P, Francois R, Wardle L, and Smith R (1999). The Iridium main mission antenna concept. *IEEE Transactions on Antennas and Propagation;* 47 (Mar.); 416–424.

Seymour D (2000). L band power amplifier solutions for the Inmarsat space segment. *IEE Seminar on Microwave and RF Power Amplifiers;* Dec. 7; 6/1–6/6.

Tanaka M, Yamamoto K, Egami S, and Ohkubo K (1996). S-band multiple beam transmitter for N-Star. *AIAA International Communications Satellite Systems Conference;* Pt. 1 (Feb. 25–29); 80–85.

Tesat-Spacecom, GmbH, & Co (2007). Facts, products, and services—active RF products. Fact sheet. On www.tesat.de/images/stories/PDF/tesat_active_products.pdf. Accessed Jul. 8, 2011.

Thales Alenia Space (2005). Channel amplifier/linearizer. Product datasheet. Feb. On www.thalesgroup.com/Pages/Solution.aspx?id=3435&pid=1500. Accessed Jan. 27, 2011.

Thales Alenia Space (2006). C, Ku, Ka bands channel amplifiers. Product datasheet. Nov. On www.thalesgroup.com/Pages/Solution.aspx?id=3435&pid=1500. Accessed Jan. 27, 2011.

Thales Electron Devices (2001). Space, advanced technologies for peak performance. Product brochure. May. Received from TED Ulm in 2006 May.

Thales Electron Devices (2009). Space K-Ka band TWT; Space Ku-band TWT; Space C-band TWT; Space L-S band TWT; Space X-band TWT. Product literature. On www.thalesgroup.com/Markets/Security/What_we_do/Radio_Frequency_and_Microwave_sources/RF_and_Microwave_sources_for_space/Space_tubes_and_amplifiers/. Accessed Jan. 27, 2011.

Villemazet J-F, Yahi H, Lopez D, Perrel M, Maynard J, and Cazaux J-L (2010). High accuracy wide band analog predistortion linearizer for telecom satellite transmit section. *IEEE MTT-S International Microwave Symposium Digest;* May 23–28; 660–663.

Weekley JM and Mangus BJ (2004). TWTA versus SSPA; a comparison of on-orbit reliability data. *IEEE International Vacuum Electronics Conference;* Apr. 27–29; 263.

Weekley JM and Mangus BJ (2005). TWTA versus SSPA: a comparison of on-orbit reliability data. *IEEE Transactions on Electron Devices;* 52 (May); 650–652.

Yuen CH, Yang SS, Adams MD, Laursen KG, inventors; Space Systems/Loral, Inc, assignee (1999). Broadband linearizer for power amplifiers. U.S. patent 5,966,049. Oct. 12.

Zhang W-M and Yuen C (1998). A broadband linearizer for Ka-band satellite communication. *IEEE MTT-S Microwave Symposium Digest;* 3 (June 7–12); 1203–1206.

CHAPTER 7

PAYLOAD'S COMMUNICATIONS PARAMETERS

7.1 INTRODUCTION

This chapter presents the communications parameters found in the top-level payload requirements document with the intent to fully explain the following for each:

- What the parameter is
- How the parameter's application at unit level relates to its application at payload level
- How the parameter is verified.

What these parameters mean to the signal quality is discussed in Chapter 10, after the basics of communications theory are presented there.

Figure 7.1 gives the simplified transponder diagram used in the sections below, for marking the payload elements that contribute to a parameter value, when there is more than one such element. ("Element" means either a unit or an element used to integrate the payload, such as a piece of waveguide.) The high-power amplifier (HPA) subsystem is identified in the figure because reference is made to it in a couple of places.

Each payload element enhances the reception and retransmission of the signal in some way but also inevitably distorts the signal in some minor way. The elements can be divided into two classes: those that cause basically linear distortion and those that cause nonlinear distortion. The effect of linear distortion

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

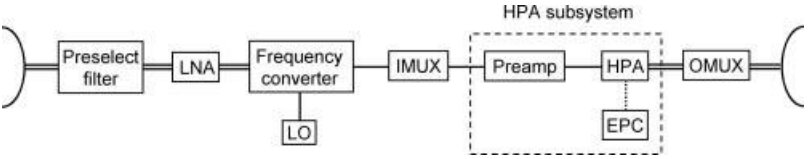


FIGURE 7.1 Simplified transponder diagram for showing main contributors to each communications parameter.

on communications is easier to analyze. The elements that are significantly nonlinear are the following:

- Frequency converter
- High-power amplifier (HPA), especially spacecraft edges and corners that cause passive intermodulation products.

Not all payload parameters represent distortions, but for those that do, a recommendation is made in each section on how to combine the signal distortion as measured or estimated for every payload element into a distortion value at payload level. We emphasize the combining upward to payload level instead of the allocation downward from payload level to the elements because typically the payload engineer starts at the beginning of a satellite program with a collection of preliminary element specifications from an earlier similar payload and must check if they will collectively fit together for the new payload. During the course of the program, the payload engineer continually refines his calculations to show that he continues to expect to meet the payload-level requirements. For some types of distortion there is no way to do the combining that makes perfect sense. In that case, a sensible way that is readily computable is given. There are special cases, not addressed in this chapter, for which the payload engineer must take another approach to distortion combining that he himself must develop. The recommended method for combining unit-level distortions sometimes depends on the channel bandwidth.

Some parameter descriptions hinge on the specification bandwidth of applicable payload elements. Bandwidth is characterized by the percentage it represents of the element's center frequency: narrow, on the order of 0.5%; wide, on the order of 3%; and very wide, on the order of 15%. For such parameters we consider no wider bandwidths, which are uncommon.

For every communications parameter, the **verification** method will be briefly presented. Verification shows that a specification is met. The customer and the payload engineer must come to an agreement on the "verification matrix," which is a table that shows the verification method proposed for every parameter in the payload requirements document. The most common method and the one regarded as the most robust, although it has its own problems, is direct test or measurement. However, for some parameters test is not necessary: parameters that were measured at a lower level of integration and that are truly not affected

by the further level(s) of integration. In this case the parameter value is "carried up" from the lower level of test. For some parameters a direct measurement is not possible, in which case the method "analysis" is applied. The analysis is based on other parameters that were directly measured. The two other verification methods, "inspection" and "by design," do not apply to the communications parameters we discuss in this book.

The parameters must meet specifications over the life of the payload. There are various reasons that some parameters change values over life. Some do because a unit's performance is temperature-sensitive. The temperature of most of the payload equipment generally rises over on-orbit life in a known way. So the lifetime parameter variation from temperature can be calculated from the temperatures over life and the beginning-of-life unit temperature-sensitivity. Other parameter values change because a unit's or component's performance slowly changes simply from being operated. The payload manufacturer relies on long-term test results on similar devices from the unit or component manufacturer. Other parameter values change because a unit or component is sensitive to accumulated radiation exposure on orbit, and here again the data comes from the manufacturer. If the sensitive device is not a unit but is incorporated into a unit by the payload manufacturer, the unit engineer uses the data to make a prediction for the unit. For more general information on effects, see Section 2.3; for more information specific to a unit, see Chapters 3–6.

The rest of the sections in this chapter are for the following communications parameters:

- *Section 7.2*: Gain variation with frequency
- *Section 7.3*: Phase variation with frequency
- *Section 7.4*: Channel bandwidth
- *Section 7.5*: Phase noise
- *Section 7.6*: Frequency stability
- *Section 7.7*: Spurious signals from frequency converter
- *Section 7.8*: High-power amplifier nonlinearity
- *Section 7.9*: Spurious signals from high-power amplifier subsystem
- *Section 7.10*: Stability of gain and power-out of high-power amplifier subsystem
- *Section 7.11*: Equivalent isotropically radiated power (EIRP)
- *Section 7.12*: Figure of merit of the payload in its function as a receiver, gain over system noise temperature (G/T_s)
- *Section 7.13*: Self-interference
- *Section 7.14*: Passive intermodulation products (PIMs)
- *Appendix 7.A*
  - *Section 7.A.1*: Antenna testing
  - *Section 7.A.2*: Relation of gain and phase ripple
  - *Section 7.A.3*: Independence of G/T_s on reference location.

7.2 GAIN VARIATION WITH FREQUENCY

7.2.1 What Gain Variation with Frequency Is

Every two-port device can be considered as a filter in that it can be characterized by a transfer function $H(f)$, where f is frequency in Hz (for more details see Chapter 10). For an active device the transfer function depends on the RF drive level input to it and the component-amplifier biases. A transfer function has two parts, the gain response $G(f)$ and the phase response $\varphi(f)$:

$$H(f) = 10^{G(f)/20} e^{j\varphi(f)}$$

where $G(f)$ is in dB and $\varphi(f)$ is in radians.

Payload or unit **gain variation with frequency** has a few names but the main ones are **gain flatness**, **gain tilt**, and **gain ripple**. Putting values to them is best done with the qualifier "pk-pk," "pk-to-pk," or "p-p," meaning "peak-to-peak," that is, highest relative to lowest in dB, which is unambiguous. (Examples of ambiguous qualifiers are "maximum" and no qualifier at all.) Gain flatness is simply the highest gain in dB over the specified band minus the lowest gain over the band, as illustrated in Figure 7.2. When gain flatness is specified, no additional gain-variation parameters are specified. An alternative characterization of gain variation is gain tilt and gain ripple together. Gain tilt is the delta gain over the band of the best-fit line over the band; it can be positive or negative. Gain ripple is then the variation in the gain over the band that is left after the tilt is taken out. Gain tilt and ripple are illustrated in Figure 7.3. This example is in fact the decomposition of the simple gain flatness of Figure 7.2.

When the band has ripple that is more or less a sinewave over frequency, the ripple slope may additionally be specified. The slope together with the ripple amplitude give you the ripple frequency period ($=\pi$ times pk-pk ripple in deg divided by ripple slope in deg/Hz). The sinewave may be in any phase.

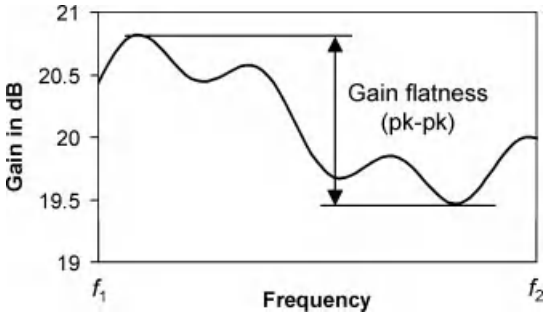


FIGURE 7.2 Definition of gain flatness.

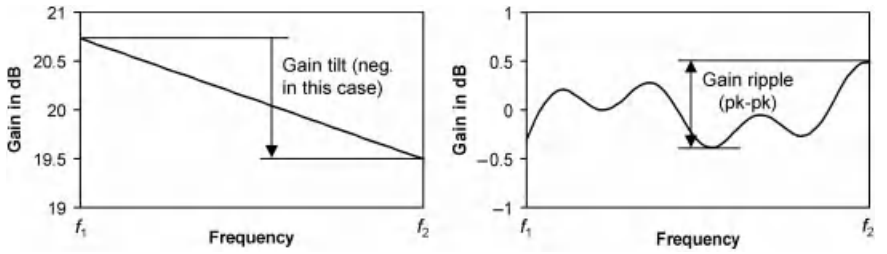


FIGURE 7.3 Definitions of gain tilt and gain ripple.

7.2.2 Where Gain Variation with Frequency Comes From

We look at the gain variation characteristics of the payload units each over their own specified band. The following is meant to be a guide to expectations of gain-variation specification and test data results but may not be perfectly accurate in all cases.

Most units have so little gain variation that they are specified on simply gain flatness, as shown in Figure 7.4. This is true for the filters (preselect, IMUX, and OMUX) because they are equiripple or nearly equiripple (Section 4.5.1.2). It is true of the frequency converter and the SSPA.

The complementary set of payload elements has gain tilt, as shown in Figure 7.5. First are the antennas. The most common types of communications-satellite antennas for the GHz frequencies are the reflector antenna and the phased array. Their apparent area measured in square wavelengths goes up with higher frequency. A reflector antenna's gain goes up linearly with frequency, in dB. For phased arrays, it is usually somewhat less than that because the common use of phase shifters instead of delay elements causes the beam to point a little off-axis for frequencies above and below mid-band (beam squinting) (Chang, 1989). Besides the antennas, waveguide, coax, the (L)CAMP, and the TWT have gain tilt, too.

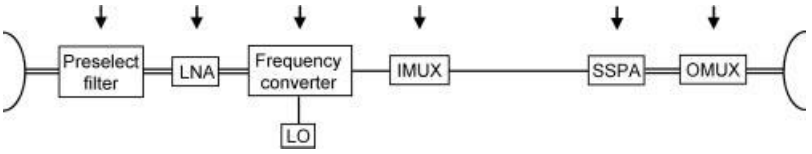


FIGURE 7.4 Payload elements whose gain variation is specified on simply gain flatness.

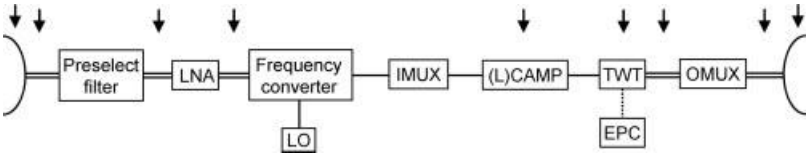


FIGURE 7.5 Payload elements with gain tilt.

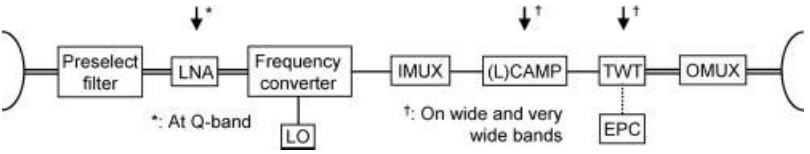


FIGURE 7.6 Payload elements with quadratic gain component.

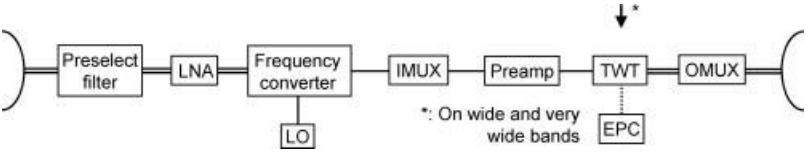


FIGURE 7.7 Payload element that often has sinewave-ripple gain component.

The (L)CAMP and the TWT have a quadratic component of gain variation in addition to tilt, on wide and very wide bands. Additionally, a Q-band LNA has been seen to have a quadratic component (Grundbacher et al., 2004). These three units are marked in Figure 7.6.

The TWT, marked in Figure 7.7, is the only unit that may have a sinewave-ripple versus frequency component of gain variation. It arises from multipath in the TWT, principally from reflections at the output port and the mid-TWT attenuator (L-3, 2009). Current TWT literature shows the ripple period to be roughly 2–3% of the center frequency. There may be one ripple, more than one ripple, or no ripple. The TWT gain variation is measured at both small-signal and saturation. The ripple amplitude at small-signal is several times what it is at saturation (L-3, 2009).

7.2.3 How Gain Variations with Frequency at Unit Level Carry to Payload Level

If the payload carries more than one channel then at the payload level the gain variation specifications will be on two layers. One layer is by the channel, to insure minimal distortion of the channel. Ideally, all the channels have the same specification if they will all carry the same kind of modulation; however, one or two channels may need different specifications if they are near another signal such as a command channel and need to be protected from it and to protect it. The second layer of payload specifications will typically be over a bandwidth wide enough to include all the channels, to insure that the channels receive approximately equal treatment.

For the payload elements whose gain variation is specified simply on pk-pk gain flatness, it is difficult to say how to combine them into a payload-level estimate. If the gain variation plot of each element across the band reminds you of plots with jumpy measurement error then it is probably best to rss together the rms values of the elements. **Rss** is the square root of a sum of squared values, and **rms** is the

square root of the average of a set of squared values. Taking the rms value across a band minimizes the impact of any large but narrow deviation, which is a legitimate approach if the signal going through the band will fill the band. Taking the rss as a way to combine all the elements' values is a good idea if the gain variations from element to element look unrelated, for example, do not all have a peak at the same frequency. The last step would be to multiply the rss value by $\sqrt{2}$ to create an approximation of a pk-pk value. There are many cases in which this is *not* a good approach. The payload engineer must inspect the gain variation plots of the major contributing elements to be able to form an idea of what is a good way to combine. For example, if two elements have a large variation at about the same frequency then the best thing to do is to form the algebraic sum of the large variations and add it to the $\sqrt{2}$ times the rss of the rms values of the other elements. For another example, if two elements have a large variation at different frequencies then the best thing to do is to take the larger of the two variations and add it to $\sqrt{2}$ times the rss of the rms values of the other elements.

For payload elements whose gain variation specs are on tilt, if consistently signed the gain tilts over a band can be added to obtain the total tilt through the payload on that band. The same goes for quadratic gains.

7.2.4 How to Verify Gain Variation with Frequency

A scalar network analyzer is used to measure the payload's end-to-end gain variation with frequency.

When the HPA's specification bandwidth is at least wide, two measurements of the HPA must be made. One is the **small-signal frequency-sweep**, for which the sweeping CW is at a constant level, low enough for the HPA to be operating in its quasi-linear region across the full band. The other is the **large-signal or saturated frequency sweep**, for which the sweeping CW is also at a constant level, high enough to saturate the TWTA across the full band. Both sweeps are required in order to provide enough information if the HPA must be modeled in a simulation (Section 13.3.6).

7.3 PHASE VARIATION WITH FREQUENCY

7.3.1 What Phase Variation with Frequency Is

Phase response $\varphi(f)$, in radians, is a frequency-domain characterization. Phase response is more useful than **group-delay** response for digital communications because its effect on the signal phasors is easier to visualize. If group delay is provided, the phase can be obtained by an integration because group delay is $-d\varphi(f)/(df/2\pi)$.

Payload or unit **phase variation with frequency** has a few names, but the main ones are **phase linearity**, **quadratic phase**, and **phase ripple**. Putting values to them is best done with the qualifier "pk-pk" because it is unambiguous. The phase response is assessed after its slope is removed, since the slope corresponds only to

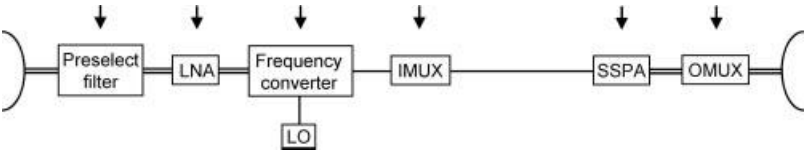


FIGURE 7.8 Payload elements whose phase variation is specified on simply phase linearity.

the average delay through the payload or unit. Phase linearity is actually deviation from linear.

7.3.2 Where Phase Variation with Frequency Comes From

We look at the phase variation characteristics of the payload units each over their own specified band. The following is meant to be a guide to expectations of phase-variation specification and test data results but may not be perfectly accurate in all cases.

Most units have so little phase variation that they are specified on simply phase linearity, as shown in Figure 7.8. This is true for the filters (preselect, IMUX, and OMUX) because they are equiripple or nearly equiripple (Section 4.5.1.2). It is true of the frequency converter and the SSPA.

Most of the rest of the payload elements, but not the antennas, have quadratic phase, as shown in Figure 7.9. The only units that do are the (L)CAMP and the TWT. Besides them, waveguide does.

The (L)CAMP and the TWT have a cubic component of phase variation in addition to quadratic, on wide and very wide bands, as shown in Figure 7.10.

The TWT, marked in Figure 7.11, is the only unit that may have a sinewave-ripple versus frequency component of phase variation. It arises for the same reason that periodic gain ripple does (Section 7.2.2) and has the same characteristics.

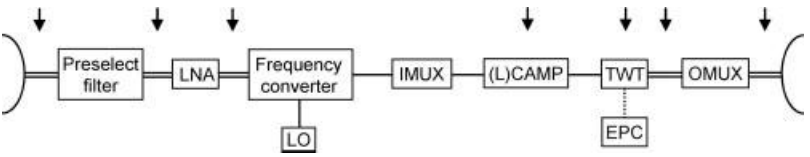


FIGURE 7.9 Payload elements with quadratic phase component.

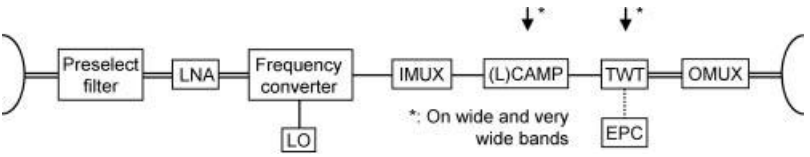


FIGURE 7.10 Payload elements with cubic phase component.

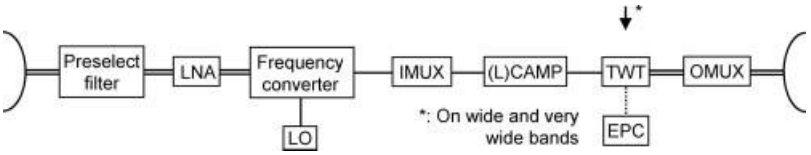


FIGURE 7.11 Payload element that often has a sinewave-ripple phase component.

Because the ripple source is passive, the phase ripple is directly related to the gain ripple; there is 7° pk-pk of phase ripple for every 1 dB of pk-pk gain ripple, and the phase and gain ripples are 1/4 cycle apart in the frequency domain. This is shown in Appendix 7.A.2.

7.3.3 How Phase Variations with Frequency at Unit Level Carry to Payload Level

Phase variations with frequency at unit level carry to payload level in the same way that gain variations do (described in Section 7.2.3). The only difference is that instead of tilts being summable, if consistently signed, quadratic and cubic components are.

7.3.4 How to Verify Phase Variation with Frequency

There is more than one way to measure the repeater's end-to-end phase variation with frequency. Because of the frequency converter in the payload, any measurement technique requires either that the LO signal be accessible or that its signal be replaceable by an external LO, because the same LO must feed both the payload and a reference mixer. Any phase measurement uses a vector network analyzer. The best way to measure the phase variation may involve a reference mixer and a reciprocal calibration mixer, where a reciprocal mixer is one that has the same gain and phase response in both up- and downconversion (Agilent, 2003).

When the HPA's specification bandwidth is at least wide, two phase measurements of the HPA must be made, same as for gain response (Section 7.2.4).

7.4 CHANNEL BANDWIDTH

The term **channel bandwidth** can be used in two ways. The more common one, which we do not mean here, is as the bandwidth over which some requirement must hold. This is better called the "specification bandwidth." For us, if a certain channel bandwidth is a requirement, it is a replacement requirement for gain and phase responses. A typical type of bandwidth specified is the **3-dB bandwidth**, which is defined in Chapter 10. See that chapter also for more background information on filters and bandwidths. We mean channel bandwidth to be the bandwidth of the channel on its total path through the payload.

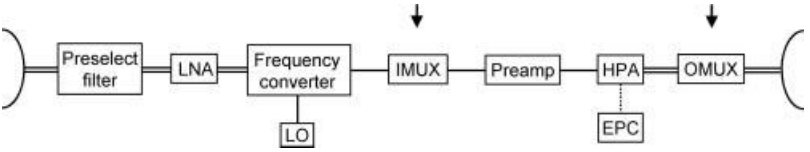


FIGURE 7.12 Sources of channel-bandwidth limitation.

The units that primarily determine a channel's bandwidth are marked with arrows in Figure 7.12. They are the IMUX and to a lesser extent the OMUX (Section 4.5.3.2).

The gain response over the channel's frequency band, through the whole payload, must be characterized at cold operating temperature and again at hot operating temperature. The intersection, in the sense of set theory, of the two responses is taken to show specification compliance. When the filter is made of waveguide, the filter response shifts from air to vacuum, so measurements in air must be adjusted accordingly (Section 4.3.2.3).

7.5 PHASE NOISE

Phase noise is an unwanted variation in the signal phase that comes from the local oscillators used in frequency conversion (Section 5.5.3). See Section 10.5.2 for a description of what part of the phase noise spectrum should have specifications on it and how phase noise affects the signal.

There are three types of phase noise test set currently in use. The simplest one is a spectrum analyzer augmented with phase noise-testing software. This type does not differentiate between phase noise and amplitude noise but combines them. However, apparently amplitude noise is minimal in an oscillator (e.g., Anritsu (2008)). Such a test set measures the single-sideband phase noise L down to 10 Hz at the lowest extreme. A more complex and better performing type of test set is the signal source analyzer (e.g., Agilent (2007)). It is also based on a spectrum analyzer but is much more. It is able to separate out phase noise from amplitude noise. It measures phase noise in two or three ways including a cross-correlation technique. It can measure L down to at least as low as 1 Hz (you do not need it this low, but this helps you to identify the kind of test set you have). A third type of test set is all digital and can measure phase noise at frequencies down to 0.1 mHz but currently only for oscillators with frequency up to 30 MHz, so they are only usable for the payload's master oscillator.

With whichever type of test set is used, the engineer must make certain that if the test set itself adds phase noise, its spectrum at all frequencies of interest is much smaller than that of the device under test.

7.6 FREQUENCY STABILITY

The **frequency stability** of the LO, if it derives from a reference oscillator, is the same as that of the reference oscillator in parts per million (ppm).

An oscillator is normally specified on both its long-term and short-term stability. **Long-term stability** or **drift** characterizes frequency variations over intervals from days to years. For communications, the interval is usually a year. This is systematic drift due to aging and is stated in ppm per the time interval. The engineer does not test this but relies on the numbers from the manufacturer. **Short-term stability** is stability over an interval of a few seconds at most. The oscillator's phase noise spectrum (see Section 5.5.2) describes the short-term stability on intervals of 1 s or less (1 Hz on up). For communications, stability over a few seconds (i.e., down to 0.1 Hz) does not matter if it is not extreme. The specification applies at every operating temperature and at every time in the life.

7.7 SPURIOUS SIGNALS FROM FREQUENCY CONVERTER

The mixer in the frequency converter can create four kinds of **spurious signals** (also called "spurious outputs," "spurious," or just "spurs") present in its output:

- Oscillator harmonics
- Harmonics of the input signal
- Cross-products of the input signal and LO carrier besides the intended cross-product
- Leakage of input signal and LO directly through to the output.

The first three are intermodulation products.

What the payload engineer can do to mitigate the spurious signals is the following:

- Ensure that the satellite's frequency plan keeps the spurious signals of significant power out of the channels
- Specify low levels of IMPs for the frequency converter
- Add filters where required.

The spurious specification applies at all operating signal levels. The specification depends on signal level because there are some spurious signals, namely those with a contribution from the signal, whose level is different at a different signal level.

When any signal input to the payload is only narrow- or wide-band but not very wide-band, there is usually no difficulty in dealing with the spurious outputs. That is, the spurious signals are so far out of the band of the downconverted signals that the only concern is interference to signals from other systems being received on the ground. However, it may happen with a very wideband signal that the spurious outputs would be a problem with only one frequency conversion on the satellite, in which case a dual conversion must be performed.

If the level of any of the spurious signals would be too high at the payload output given the passive filtering already foreseen and the filtering provided by the active

units then a passive filter must be added, carrying a penalty in RF loss, mass, and spacecraft-panel mounting area.

The spurious signals are measured with a spectrum analyzer by sweeping a CW at the frequencies where the spurious signals may be. A CW is used because it is easier to see the mixer outputs than with a modulated signal. The type of spurious signal seen can be partially found out by inspection of whether and how it moves as the CW is swept, since the mixer output (m,n) will move m times as fast as the CW is moving and in the direction given by the sign of m. A report of a test result must identify the resolution bandwidth used, as this is the bandwidth over which the analyzer averages power. A small spurious signal can be visible with a small resolution bandwidth but not be noticeable with a larger bandwidth. The resolution bandwidth is part of the spurious specification.

7.8 HIGH-POWER AMPLIFIER NONLINEARITY

What the HPA nonlinearity is was discussed in Section 6.2. The four ways in which the nonlinearity can be specified were also discussed in that section and are as follows:

- P_{out} versus P_{in} curve and phase shift ϕ versus P_{in} curve
- Derivative of P_{out} versus P_{in} curve and derivative of phase shift versus P_{in} curve
- C/3IM
- NPR.

Verification of the HPA nonlinearity depends on the way it is specified. To measure P_{out} versus P_{in} and ϕ versus P_{in}, the vector network analyzer is used, but instead of sweeping the frequency, you sweep the signal level. The CW stays at one frequency during the power sweep. If the HPA is to be used on a very wide band, the measurements should be taken at about three frequencies, at midband and at about 1/6 and 5/6 of the way across the band. Verification for the other specification methods were described in Section 6.2.2.

The nonlinearity is measured at HPA level and carried up to payload level.

7.9 SPURIOUS SIGNALS FROM HIGH-POWER AMPLIFIER SUBSYSTEM

7.9.1 What HPA-Subsystem Spurious Signals Are

A HPA puts out spurious signals near the carrier. They are better described in the literature for the TWTA than for the SSPA.

For a TWTA, some of them exist whether or not a signal is present, and some exist only when there is a signal (L-3, 2009). Their frequencies are usually between 100 Hz and 500 KHz from the carrier (L-3, 2009). These spurious signals are so close to the carrier frequency that they cannot be filtered out. A TWTA's spurious signals arise from both **spurious phase modulation (PM)** and **spurious**

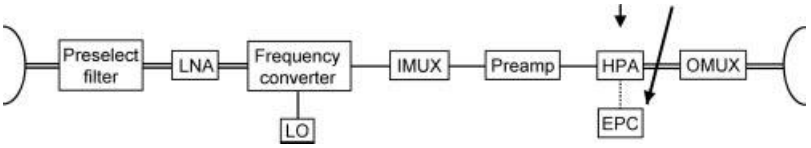


FIGURE 7.13 Sources of HPA-subsystem spurious signals.

amplitude modulation (AM). A spurious PM that has a given frequency can be represented by a small sinewave component of the signal phase. A spurious AM that has a given frequency can be represented by a small sinewave component of the signal amplitude. On a spectrum analyzer one of either type would be seen to have many harmonics, symmetric in spacing and power about the carrier. On a phase noise test set, spurious PM would be seen as a delta-function on the phase noise spectrum.

The spurious signals specification applies at every operating signal level. There are some spurious signals, namely those with a contribution from the signal, whose level is different at a different signal level.

7.9.2 Where HPA-Subsystem Spurious Signals Come From

For the TWTA, both the TWT and the EPC create the spurious signals, and for a SSPA it may be the same, as illustrated in Figure 7.13.

For a SSPA, there are in-band spurious and EPC switching-noise spurious (AIAA, 2003).

For a TWTA, the cause of the spurious signals that exist even without a carrier present is usually ripple on power supply voltages:

- The cathode heater, usually a filament, causes PM if it is connected to an alternating-current (AC) power supply.
- Cathode voltage fluctuation in the form of ripple causes PM and AM, but the spurious signals due to AM are nearly always 10 dB lower than those due to PM. The ripple has two sources: AC components at the fundamental and harmonics of the DC-to-AC converter's chopper (which precedes the voltage transformer and the AC-to-DC converter in the power supply) and AC ripple or transients on the power supply's input bus voltage. A change in cathode voltage affects the speed of the electrons in the beam, which affects the phase of the RF output signal.

The cause of the spurious signals that exist only when a carrier is present is interactions within the TWT itself:

- Cathode voltage fluctuation at a rate equal to the modulation-symbol rate R_s causes PM. Since the RF signal being amplified has variations in its signal level, the amount of electron current that undesirably hits the TWT's helix also varies, which causes the cathode voltage to fluctuate.

The source for the material in this paragraph is L-3 (2009).

7.9.3 How HPA-Subsystem Spurious Signals Carry to Payload Level

Spurious that could interfere into other channels may be removed by the OMUX. For the TWTA, since the maximum frequency of these spurious signals is 500 KHz, for this to apply the carrier spacing would have to be less than 1 MHz, which is low for communications satellites. Spurious signals due to PM do not harm the signal once the signal is received on the ground (Section 10.5.2). The spurious that are due to AM will degrade the BER unless the modulation symbol rate is much less than 500 KHz. Also, if the antenna beam that carries this signal impinges on another beam, this signal will cause a varying level of interference that must be accounted for in the self-interference calculation (Section 7.13). For spurious PM due to signal-amplitude fluctuation, they will be mostly or entirely removed by the detection filter on the ground (Section 10.5.3) so will cause little or no harm.

7.9.4 How to Verify HPA-Subsystem Spurious Signals

At both the unit level and payload level, these spurious signals can be measured with a spectrum analyzer with a CW signal. However, if it is desired to determine which spurious signals are due to PM and which are due to AM, two special test setups are required (L-3, 2009).

7.10 STABILITY OF GAIN AND POWER-OUT OF HIGH-POWER AMPLIFIER SUBSYSTEM

7.10.1 What Gain Stability and Power-Out Stability Are

Stability of gain and **stability of power-out** are characteristics of the gain and power-out at the input to the transmit antenna. Stability is the change in overall level of the two parameters over the operating temperature range and life, with all other conditions being equal. The stability may be specified at only the mid-channel frequency or at a few frequencies.

A particular payload may have either one of these parameters specified or both, depending on the operating modes of the preamp. If the unit has a fixed-gain mode (FGM) then the unit has the capability to amplify the signal by a commanded fixed gain. If the unit has an automatic level-control (ALC) mode then it has the capability to the amplify the signal (plus distortions, interference, and noise) to a commanded level. The commanded level corresponds to a particular HPA input backoff for noise-free single-carrier operation (Section 6.2.1). When the unit is in FGM, then the applicable specification is gain stability; when it is in ALC mode, the applicable specification is power-out stability.

7.10.2 Where Gain Stability and Power-Out Instability Come From

The stability of the gain and power-out depend primarily on characteristics of the preamp and of the HPA, as depicted in Figure 7.14. The only other possibility is the OMUX, but as long as it is on heat pipes its effect is probably negligible.

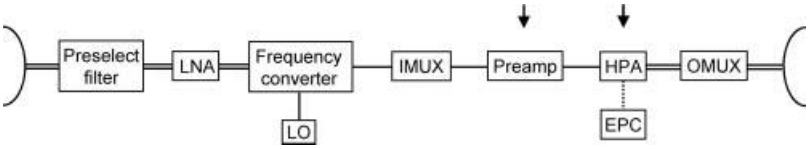


FIGURE 7.14 Sources of gain and power-out instability.

TWTA performance changes over temperature and life are mostly well characterized, and the (L)CAMP compensates them. The (L)CAMP also compensates its own changes (Section 6.4.9).

The SSPA compensates it own performance change over temperature. With aging, its gain decreases for a fixed RF drive level (Section 6.5.5).

For both types of HPA, gain and power-out instability are only residual effects from imperfect compensation.

7.10.3 How Gain Stability and Power-Out Stability Carry to Payload Level

These parameters are only defined at the payload level.

7.10.4 How to Verify Gain Stability and Power-Out Stability

The contributing factors to gain stability and power-out stability, exclusive of aging and radiation, are tested at the unit level. The parameters themselves, except for variation from aging and radiation, are tested at system test. The tendencies that the units have over life, due to aging and radiation, are known from earlier testing of such units by their manufacturers, and these effects are factored in by analysis.

7.11 EQUIVALENT ISOTROPICALLY RADIATED POWER

Equivalent isotropically radiated power (EIRP) is often the first payload parameter that comes to mind. It tells how much power the payload must radiate toward every spot in the payload's transmit coverage area, which is the combined area on the earth over which the payload's transmission must meet minimum-power and other specifications. Only for those communications satellites that also receive signals from users, is the uplink coverage area of equal significance; for that, the pertinent payload parameter is G/T_s (Section 7.12). EIRP is the product of the power into the transmitting antenna's terminal and the antenna gain in a given direction.

The main contributors to EIRP in the positive sense are the preamp, the HPA, and the transmit antenna. The preamp sets the RF drive level into the HPA, thereby determining the HPA's output power. The negative contributors are the OMUX and the post-HPA RF lines. Figure 7.15 illustrates this.

EIRP is a payload-level requirement. Actually, it is a satellite-level requirement because it depends on the antenna's on-orbit pointing error, which depends on the spacecraft bus's ability to control it. For a spot-beam payload at GEO the satellite's pointing error can be a critical factor in the EIRP.

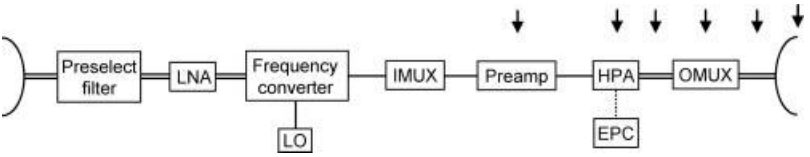


FIGURE 7.15 Payload elements that determine EIRP.

EIRP without consideration of antenna pointing error is verified in spacecraft-level testing at the compact antenna test range (CATR) (see Appendix 7.A.1 on antenna testing). The effect of pointing error on the EIRP is folded in analytically.

7.12 FIGURE OF MERIT G/T_s

7.12.1 What G/T_s Is

Part of the payload's function is to serve as a receive terminal for the uplink signal. That aspect of the payload can be characterized by the terminal figure of merit G/T_s. (It is usually written "G/T" but we want to emphasize what particular T it is and to differentiate it from T used elsewhere in this book.) G is the gain of the receive antenna and is stated in dBi, namely dB relative to the gain of an ideal isotropic antenna (with gain 0 dB). $\boldsymbol{T_s}$ is the **system noise temperature** of the payload and is stated in degrees Kelvin or usually just Kelvin. G/T_s is stated in units of dBi/K, which means G in dBi minus T_s in the form of $10\log_{10}T_s$. Both G and T_s must be **referenced** to, that is, apply at, the same point in the payload, usually the antenna terminal. (Any point can be the reference point, as is shown in Appendix 7.A.3, and the answer comes out the same.) The G/T_s definition is illustrated in Figure 7.16.

The usage of G/T_s is that the uplink SNR can be calculated from it, if the signal level into the antenna is known:

$$\text{Uplink SNR} = \frac{S}{N}$$

where

$S = P_{den}(c/f_c)1/(4\pi)G$
P_{den} = signal power density at input to antenna in units of power/m^2
c = speed of light = 2.998E8 m/s
$N = T_s\kappa B$
κ = Boltzmann's constant = 1.379E-23 W/(°K-Hz)
B = equivalent noise bandwidth on which noise is measured

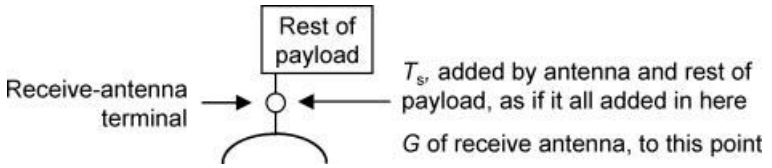


FIGURE 7.16 G/T_s definition.

The terms G and T_s require further explanation.

G is the receive antenna gain in the direction of the signal, taking the pointing error into account in some way agreed with the customer.

The system noise temperature T_s is a little more complex. It has two inband components that the payload produces and adds to the signal:

- *Antenna Noise Temperature*, T_a: The noise that the antenna receives from where it looks
- *Thermal Noise Temperature*, T_e, *due to Rest of Payload*: The noise generated in every transponder element, whether active or passive, besides the antennas

There is an excellent treatment on the noise components in Pritchard and Sciulli (1986). We have

$$T_s = T_a + T_e$$

Antenna temperature T_a is not actually due to the antenna but to what the antenna receives along with the uplink signal. See Section 11.4 for discussion on antenna temperature.

The thermal noise temperature T_e is due to the rest of the payload. It can be estimated by combining the noise figures (NF) of all the payload elements past the G/T_s reference point by the well known method represented in Figure 7.17. If F is the combined noise figure, the thermal noise temperature is given by

$$T_e = (F - 1)290 \text{ in } ^\circ\text{K}$$

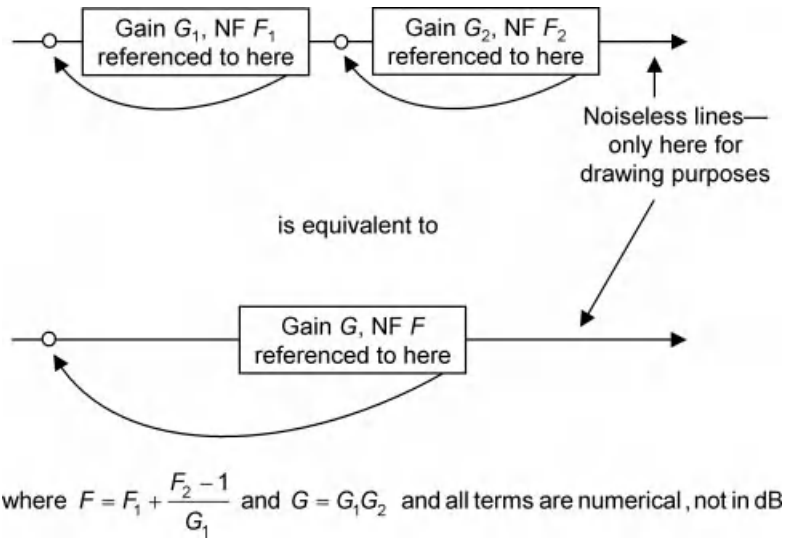


FIGURE 7.17 How to combine noise figures of consecutive elements.

The payload's noise figure merits some discussion. The payload will receive a very weak signal, which will have to go through many elements in the payload. Most of these elements cause a loss to the signal, and some like the HPA have very large noise figures. The noise figure must be **set** in the **front end** of the payload, which means that shortly past the antenna there must be a unit which primarily determines the noise figure for the entire payload. This unit is the LNA. Its requirements are that it greatly amplify the signal while adding little noise. If the payload is required to handle a large range (say 30 dB) of input signal level, the LNA may not be able to set the noise figure for all signal levels. The preamp is the unit that must compensate, fully or partially, the large range of payload input-signal levels. It may do this by amplifying the signal by many dB and then attenuating by the amount required. In this case, the preamp will have such a large noise figure that it contributes noticeably to the overall payload noise figure but is still not the main determinant of it. Alternatively, this unit may switch in the amplifier stages that it needs and then do only a small amount of attenuation. In this case, the noise figure may not be so bad. The bottom line is that the uplink SNR is satisfactory for all input signal levels because the larger noise figure only happens when the signal level is also larger. The payload's noise figure must be measured or calculated in at least two cases, that is, when the preamp is configured for minimum attenuation and for maximum attenuation. There is more discussion on dealing with noise figure in Section 8.2.

7.12.2 How to Verify G/T_s

G/T_s is verified more than once.

The antenna gain is measured for the first time on the near-field range (see Appendix 7.A.1).

The antenna temperature is a value that must be agreed upon with the customer. It is typically the worst case that the antenna will see on orbit.

The noise figure of the payload is measured during thermal-vacuum testing so the worst case over temperature can be obtained. Noise figure is measured with a spectrum analyzer augmented for noise-figure measurements, including a noise source as input. The preamp must be in fixed-gain mode. The noise figure must be measured in at least two cases, that is, when the preamp is configured for minimum attenuation and for maximum attenuation. The noise figure must be measured across the full frequency band, as it is not flat over the band. The way the noise figure is measured is the following. The payload's gain G must be already known from a prior measurement. The noise source puts out two possible noise temperatures, T_{off} and T_{on}, where $T_{on} > T_{off}$. The noise power out of the payload in a noise bandwidth B is measured with both temperatures as input. The two points are plotted with input temperature as the x-axis. A straight line is drawn between the two points and is extrapolated to input temperature of 0 K. The y-value at that point equals $(F-1)G\ 290\,B$, so F can be calculated from it. Such a plot is illustrated in Figure 7.18.

So from the antenna gain, the agreed-upon antenna temperature, and the noise figure, G/T_s can be calculated.

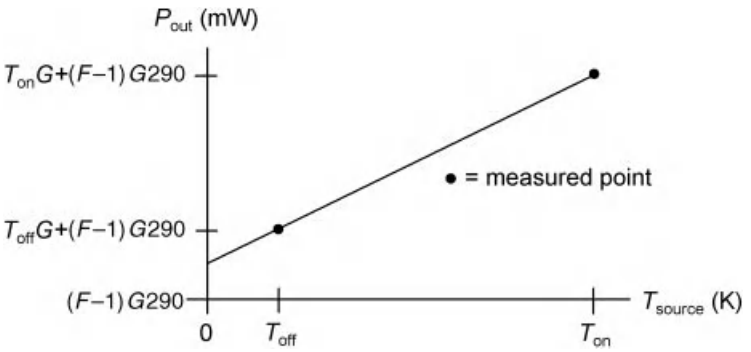


FIGURE 7.18 Plot for calculating payload noise figure.

Antenna gain is measured again in the compact antenna test range. There also G/T_s is measured.

7.13 SELF-INTERFERENCE

7.13.1 What Self-Interference Is

Self-interference is the payload interfering into itself, on transmit. It may be spot beams interfering into other spot beams or into a regional beam or a regional beam interfering into spot beams. **Carrier-to-interference ratio (*C/I*)** is the usual measure of self-interference. Since *C/I* is a function of location, practical computation limits the number of ground locations that it can be calculated for. *C/I* does not completely describe the ratio of useful signal power to distortions but is only one term in that calculation (see Section 11.5.1 for discussion on this topic).

7.13.2 Where Self-Interference Comes From

C/I has many independent dimensions, all of which must be taken into account at once:

- *Modulation on Signals in Channels*: This is a function of the uplink ground station, not the spacecraft, but *C/I* computations require knowledge of the signal spacing and signal power levels in the channel and their modulation-symbol rates and modulation schemes
- *Channel Spacing in Frequency Domain*: This depends on the size of the guard band between channels
- *HPA Nonlinearity Characteristics, as well as HPA Operating Point*: This knowledge is required for computation of the 3rd-order IMPs

- *OMUX Characteristics*: They tell how much the OMUX will reject the 3rd-order IMPs
- EIRP
- *Antenna Beam Patterns*: Spot beams do not necessarily all have the same pattern. The way the pattern rolls off is key to the computation of *I*
- Coverage requirements
- *Antenna Polarization Isolation*: This determines how much of the complementary polarization is transmitted as interference
- *Antenna Cross-Polarization*: This partially determines how much of one polarization will interfere into the other upon reception on the ground
- Assignment of channels and polarizations to the set of beams.

The *C*/*I* that is computed in this way is only one component of the total communication system-level *C*/*I*, which also includes interference from other uplink signals, downlink signals from other satellite systems, and signals from this satellite but on the other polarization.

Calculation of *C*/*I* variation requires, first of all, knowledge of the nominal *C* and *I*. There are a few extra, simultaneous dimensions that must be taken into account separately for *C* and for *I*:

- Gain stability and power-out stability (Section 7.10)
- *Antenna Pointing Error*: If the spot-beam antennas are fixed on a plate or in a structure, they will move together. On the other hand, two antennas may be separately steered. Antenna pointing error is a probabilistic phenomenon
- *Antenna Beam-Pattern Edge*: The beam pattern sharply decreases/increases at the beam edge, which is where *I* comes from.

See Chapter 12 for a probabilistic treatment of the *C*/*I* variation which is not as conservative as the usual combining of worst cases.

7.13.3 How Self-Interference Carries to Payload Level

Self-interference is a payload-level characteristic.

7.13.4 How to Verify Self-Interference

Self-interference levels are verified from a combination of testing and computation. The individual antenna beams are tested for pattern and polarization isolation, or they may be tested simultaneously so the *C*/*I* contribution from the transmit antenna can be measured (see Appendix 7.A.1). The payload channels are normally tested with CW input or inputs. Their output is taken at the test couplers immediately before the antenna terminals, for the power-out and frequency responses. The rest of the verification is by analysis.

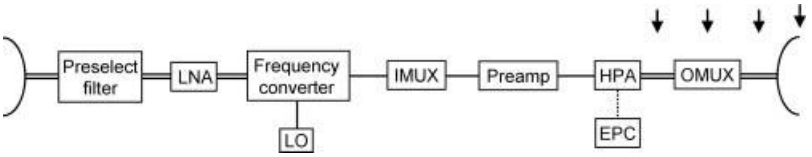


FIGURE 7.19 Sources of PIMs, besides other exposed spacecraft surfaces.

7.14 PASSIVE INTERMODULATION PRODUCTS

Passive intermodulation products (PIMs) are a phenomenon that occurs when transmitted signals hit edges or corners of spacecraft features, which perform a nonlinear combining of the signals, and then an intermodulation product happens to feed back into a receive antenna. The transponder elements that can cause PIMs are shown in Figure 7.19. If it happens that the PIM is at about the same frequency as a signal on that channel, the PIM becomes a distortion on that signal and is from then on transmitted with that signal.

When a communications satellite has many antennas and/or transmits on several frequency bands, PIMs are a significant concern. One example of a satellite for which they were a big concern is the Optus-C1 satellite, which had UHF and X-, Ku-, K-, and Ka-band payloads and 16 antennas (Singh and Hunsaker, 2004).

Satellite manufacturers have done a lot of work with PIMs, which are difficult to analyze with any accuracy so a worst-case approach is taken. The approach by one satellite manufacturer is multistage (Singh, 2008):

- First, perform analysis. It takes into account near-field antenna analysis, spacecraft structure scattering, multipath, and antenna subsystem layout
- Then assess the risk that the PIMs created will be a concern
- Design the spacecraft to avoid the occurrence of bad PIMs. Design measures can be taken on the antennas, spacecraft thruster, TWTA radiators, and spacecraft thermal blankets
- Last, test at two levels to ensure that the potentially bad PIMs are weak: at the antenna-subsystem level with a spacecraft mock-up and flight blankets and at spacecraft level including payload receivers.

Nothing further on PIMs is included in this book because they do not lend themselves to accurate analysis, aside from possible frequencies where they may occur. For further reading, see, for example, Singh and Hunsaker (2004).

APPENDIX 7.A

7.A.1 Antenna Testing

Antenna testing has a few stages. One of the first is photogrammetric measurement of the reflector surface, done in a thermal chamber (Wiktowy et al., 2003). From this data the reflector's performance over temperature can be computed.

Next, either the receive or the transmit pattern can be measured for a passive antenna, as both would be the same. The antenna is tested along with a good mockup of the parts of the spacecraft that have any possibility of causing blockage, reflections, or PIMs (Lee, 1999). Testing is often done in a **near-field range (NFR)**, which is an anechoic chamber, lined with RF absorbers. A probe performs a fine raster scan, moving in a plane in front of the aperture in the near field, measuring amplitude and phase of the electric field. The planar movement is appropriate for high-gain spacecraft antennas (Hindman and Fooshe, 1998). The near-field pattern is then transformed via the two-dimensional Fourier transform into the far-field pattern. This manner of testing is also done for planar phased arrays (Van Rensburg, 1998). In this way, gain, cross-pol, and *C*/*I* are measured.

At spacecraft-level verification, the antenna is tested mounted on the spacecraft in a far-field range or, more commonly, in a **compact antenna test range (CATR)**. A CATR emulates a far-field range by creating a plane wave, as would be in the far field, from the wave with spherical phase front that is available at a shorter distance, by means of a focusing element (Tuovinen et al., 1997). CATR measurements are not as accurate as NFR measurements. There are three kinds of CATR for measuring payload antennas, providing accurate cross-pol measurements (Fasold, 2006):

- Single-reflector compact ranges
- Dual cylindrical-parabolic ranges, having two single-curved cylindrical-parabolic reflectors
- Compensated (double-reflector) compact range, having two double-curved and compensated reflectors.

The single-reflector CATR uses an offset paraboloid (Gillespie, 2001). In the dual cylindrical-parabolic range, the first cylindrical-parabolic antenna acts as a subreflector and transforms the spherical wave into a cylindrical wave, while the second such antenna acts as the main reflector and transforms this wave into a plane wave (Gillespie, 2001).

All these CATRs seem to provide good measurements. A long-time manufacturer of the single-reflector range developed a technique to make its cross-pol measurements be as accurate as those of the dual cylindrical-parabolic range measurements, to about −48 dBi (Rose and Cook Jr., 1999). The inventor of the compensated compact range analyzed the various CATR types when they all have the same-size quiet zone (the place where the antenna under test can be positioned). He showed that co-pol performance was similar in all but cross-pol performance was best in the compensated compact range (Fasold, 2006).

CATR measurements include gain, cross-pol, *C*/*I*, EIRP, and G/T_s at ambient.

7.A.2 Relation of Gain and Phase Ripple

Multipath, for example, from an impedance mismatch, always causes a ripple in both gain and phase. For every 1 dB of pk-pk gain ripple the phase ripple is 6.6°

pk-pk, and the phase and gain ripple are 1/4 cycle apart in the frequency domain. Suppose that the impulse response of the multipath channel is this:

$$h(t) = (1 - \varepsilon)\delta(t) + \varepsilon f e^{j\lambda}\delta(t - \tau)$$

where

ε = real number of amplitude much less than 1

λ = rotation angle in radians

τ = delay of small signal relative to main signal

Then the gain response $G(f)$ and the phase response $\varphi(f)$ of the transfer function $H(f)$ are approximately given by the following:

$$G(f) \doteq \frac{20}{\ln 10}\varepsilon[\cos(2\pi f\tau) - 1]$$

$$\text{Ripple part of } G(f) \doteq \frac{20}{\ln 10}\varepsilon\cos(2\pi f\tau)$$

$$\varphi(f) \doteq \sin(2\pi f\tau)$$

and the ratio of the pk-pk phase ripple in degrees to the pk-pk gain ripple in dB can easily be calculated.

7.A.3 Independence of G/T_s on Reference Location

In Section 7.12.1, we discussed G/T_s, the figure of merit for the payload when the payload is viewed as a receive terminal. We stated that both G and T_s must be referenced to the same point, where G is the gain up to that point and T_s is the system noise temperature after that point. The receive-antenna terminal is usually taken as the reference point. We show here that in fact any point can be chosen without altering the value of G/T_s.

Figure 7.A.1 shows two cases of the reference point location, one at the antenna terminal and the other somewhere farther down the payload hardware chain. In both cases the entire payload electronics (minus the antennas) chain is represented by the cascade of two electronic elements, where each element would actually be the cascade of any number of consecutive elements in the signal path.

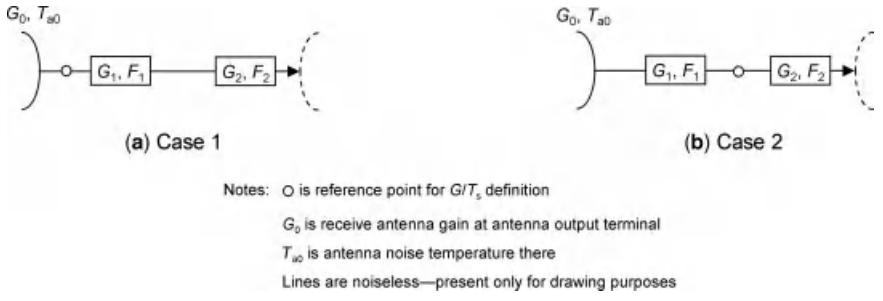


FIGURE 7.A.1 Two cases of reference point for G/T_s.

In the first of the two cases, where the reference point is at the antenna terminal, we have

Gain to reference point $G = G_0$

Electronics past reference point has combined noise figure $F = F_1 + F_2 - 1/G_1$

Therefore,

$$G/T_s = \frac{G_0}{T_{a0} + T_{r0} + (F-1)290}$$

In the second case, we have

$$\begin{aligned} G &= G_0 G_1 \\ T_a + T_r &= T_{a0} G_1 + T_{r0} G_1 (F_1 - 1) G_1 290 \\ F &= F_2 \quad \text{and} \quad T_e = (F_2 - 1)290 \end{aligned}$$

Therefore,

$$\begin{aligned} G/T_s &= \frac{G_0 G_1}{T_{a0} G_1 + T_{r0} G_1 + (F_1 - 1) G_1 290 + (F_2 - 1)290} \\ &= \frac{G_0}{T_{a0} + T_{r0} + \left[F_1 - 1 + \dfrac{F_2 - 1}{G_1} \right] 290} \end{aligned}$$

The results are the same at both locations.

REFERENCES

Agilent Technologies (2003). Agilent PNA microwave network analyzers, mixer transmission measurements using the frequency converter application. Application note 1408-1; May.

Agilent Technologies (2007). Agilent E5052B signal source analyzer, advanced phase noise and transient measurement techniques. Application note.

AIAA Japan Forum on Satellite Communications (2003). Capital products & review: C-band solid state power amplifier (SSPA). *Space Japan Review*; no. 27; Feb./Mar. On satcom.nict.go.jp/English/english2-3/index.html. Accessed June 27, 2011.

Anritsu Corp (2008). Using Anritsu's Spectrum Master™ and Economy Bench Spectrum Analyzers to measure SSB noise and jitter. Application note 11410-00461, rev A.

Chang K, editor (1989). *Handbook of Microwave and Optical Components, Vol. 1, Microwave Passive and Antenna Components*. New York: John Wiley & Sons.

Fasold D (2006). Measurement performance of basic compact range concepts. *Antenna Measurement Techniques Association Europe Symposium;* May 1–4; 1–11.

Gillespie ES (2001). A brief history of the compact range and the near-field range. *IEEE Antennas and Propagation Society International Symposium;* 4 (Jul. 8–13); 436–439.

Grundbacher R, Chou Y-C, Lai R, Ip K, Kam S, Barsky M, Hayashibara G, Leung D, Eng D, Tsai R, Nishimoto M, Block T, Liu P-H, and Oki A (2004). High performance and high reliability InP HEMT low noise amplifiers for phased-array applications. *IEEE MTT-S International Microwave Symposium Digest;* 1 (June 6–11); 157–160.

Hindman G and Fooshe DS (1998). Probe correction effects on planar, cylindrical and spherical near-field measurements. *Antenna Measurement Techniques Association Conference*.

Lee T-H (1999). Design and analysis of antenna systems properly integrated into a spacecraft structure. *IEEE Antennas and Propagation Society International Symposium;* 4 (Aug.); 2352–2355.

L-3 Communications Electron Technologies, Inc (2009). *TWT/TWTA Handbook*, 13th ed. Torrance, CA: L-3 Communications Electron Technologies, Inc.

Pritchard WL and Sciulli JA (1986). *Satellite Communication Systems Engineering*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Rose CA and Cook JH. Jr. (1999). High accuracy cross-polarization measurements using a single reflector compact range. Antenna Measurement Techniques Association Conference; Oct 4–8; 1–12. On www.mi-technologies.com. Accessed Jan. 1999.

Singh R (2008). Passive intermodulation (PIM) requirements for communications satellites. *International Workshop on Multipactor, Corona and Passive Intermodulation in Space RF Hardware;* Sept. 24–26.

Singh R and Hunsaker E (2004). PIM risk assessment and mitigation in communications satellites. *AIAA International Communications Satellite Systems Conference;* May 9–14; 1–17.

Tuovinen J, Vasara A, and Räisänen A, inventors (1997). Compact antenna test range. U.S. patent 5,670,965. Sept. 23.

Van Rensburg DJ (1998). Phased-array simulation for antenna test range design. *Antenna Measurement Techniques Association Conference*.

Wiktowy M, O'Grady M, Atkins G, and Singhal R (2003). Photogrammetric distortion measurements of antennas in a thermal-vacuum environment. *Canadian Aeronautics and Space Journal;* 49 (June); 65–71.

CHAPTER 8

MORE ANALYSES FOR PAYLOAD DEVELOPMENT

8.1 INTRODUCTION

This chapter presents various analyses, additional to those in the previous chapter, which are useful for all stages of payload development, namely specification, design, and measurement. The analyses in this chapter involve no other element of the end-to-end communications system besides the payload (those that do are presented in Chapters 10–13). Any signal is assumed to be a carrier (CW), to reflect the manner in which payloads are usually specified. Each analysis is useful alone or together with others for a higher-level analysis.

The general topics covered in this chapter are the following:

- *Section 8.2*: How to deal with noise figure (every payload engineer should be able to estimate the noise figure of the payload in his head!)
- *Section 8.3*: How to keep payload performance budgets, an all-important topic, including an in-depth exploration of how to handle uncertainty
- *Section 8.4*: Miscellaneous analyses
  - How to characterize nonlinearity of low- and medium-power amplifier components
  - How to know if HPA nonlinearity should be specified on C/3IM or NPR
  - How to ease payload integration of combined TWTAs
  - How to avoid most Monte Carlo simulations

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

- *Appendix 8.A*: Elements of probability theory for payload analysis (and for end-to-end communications system analysis of Chapters 10–13).

8.2 HOW TO DEAL WITH NOISE FIGURE

In Section 7.12, there was some discussion on the payload's noise figure (NF) in the context of system noise temperature. The present section expands on the topic. NF is defined and the formula is provided for computing the NF of a cascade of two electronic elements. Tips are given for dealing with passive elements.

8.2.1 Defining Noise Figure

At the output of every electronic element there is noise, which comes from two sources. The first source is the noise input to the element. If a resistor at temperature T_0 serves as the input to an electronic element and the resistor's output resistance and the element's input resistance are matched, then the temperature at input and the corresponding temperature at output of the element are the following:

Resistive source's temperature at element input $= T_0$

Corresponding noise temperature at element output $= T_{0\,\text{out}} = GT_0$

where G is the element gain, namely, the gain that a signal sees as it goes through the element, not in dB, and T_0 is in K. G is always positive. The input temperature has been multiplied by the element's gain. The second source of noise at element output is the thermal noise generated internally by the element itself. Then NF F is defined as the ratio of the combined output-noise temperature to the output-noise temperature due only to input noise, when $T_0 = 290$:

$$\text{Noise figure } F = \frac{T_{0\,\text{out}} + T_{\text{int}}}{T_{0\,\text{out}}} \text{ when } T_0 = 290 \text{ K}$$
$$= \frac{G\,290 + T_{\text{int}}}{G\,290} = 1 + \frac{T_{\text{int}}}{G\,290}$$

(NF is always defined on the basis of 290° input, to make the definition independent of the noise temperature actually input to the element. See Section 7.12 for how to calculate actual input temperature.) F is always greater than 1. The contribution to the output-noise temperature T_{int} that the element generates internally is related to F by the following:

$$T_{\text{int}} = (F - 1)G290$$

For the general input noise temperature T_0, the total output-noise temperature is $GT_0 + T_{\text{int}}$.

8.2.2 Noise Temperature at Input and Output of Passive Element

It is handy to know some properties of noise temperature at the output of a passive element.

A passive element has gain $G < 1$; its NF is $1/G$. If the input-noise temperature is 290, as in a NF measurement, then the output-noise temperature is also 290. This is because the input noise sees the element gain, so its temperature at output is $G290$, and the internally generated noise temperature at output is $(1/G - 1)G290 = (1 - \mathrm{G})290$, and their sum is 290. More generally, for any input-noise temperature of T_0, the output-noise temperature is $GT_0 + (1 - G)290$.

A passive element's output-noise temperature can be lower than the input-noise temperature; this happens whenever $T_0 > 290$. There is a floor, that is, minimum, to the output-noise temperature, namely 290, since when 290 is input, 290 is also output.

8.2.3 Gain and Noise Figure of Two-Element Cascade

Key to calculating NF of the payload are the formulas for combining the gains and NFs of two electronic elements in cascade, given in Figure 8.1.

It is especially easy to calculate the NF of the cascade when the first element is passive. Suppose that the passive element has gain $G_1 < 1$; then its NF is $F_1 = 1/G_1$. Suppose that the active element has NF F_2. Then the cascade's NF is simply $(1/G_1)F_2$, as illustrated in Figure 8.2.

8.2.4 Playing Off Gains and Attenuations

In this subsection, we address two questions in payload (or even unit) design that affect overall NF. Since NF is always to be minimized, good design does that. Now, amplifiers typically have a fixed gain selectable from a few values, while attenuators come in many varieties: fixed selectable from a fine-grained set of values, variable but with value set in payload integration, variable set autonomously onboard in an automatic level-control (ALC) unit, and variable commandable from the ground. So usually when an amplifier is needed, it is not quite the right value, so an attenuator must be paired with it.

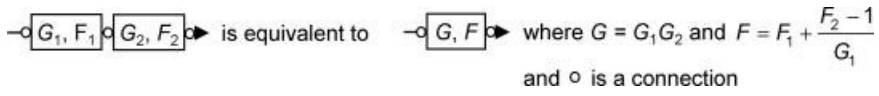


FIGURE 8.1 Gain and noise figure of cascade of two general electronic elements.

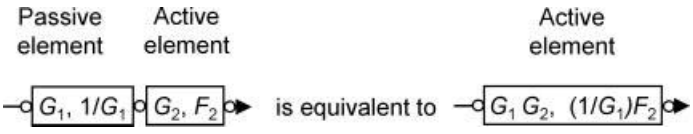


FIGURE 8.2 Gain and noise figure of passive element followed by active element.

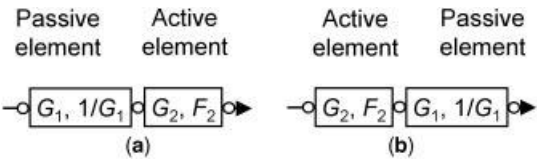


FIGURE 8.3 Same gain attained by means of (a) attenuator followed by amplifier and (b) amplifier followed by attenuator.

The first question is which is better design, to follow the amplifier with the attenuator or the attenuator with the amplifier. These two cases are depicted in Figure 8.3, where $G_1 < 1$. In the first case, the combined NF F is given by

$$F = \frac{1}{G_1} F_2$$

In the second case it is given by

$$F = F_2 + \frac{\dfrac{1}{G_1} - 1}{G_2}$$

As an example, let us assume that $G_1 = 1/\sqrt{10}$ and $G_2 = 10$. In the first case $F \approx 3.2F_2$, and in the second case $F \approx F_2 + 0.2$. In the first case F is larger, that is, worse, since $F_2 > 1$. The first case is always worse; the second design is always better.

The second question regards the HPA's preamp, which must be able to handle a large input-signal range, often about 30 dB. Recall that the preamp outputs a signal almost anywhere in a large output range independent of where the input signal lies in a large input range. The preamp could be designed to amplify by many dB and then attenuate as much as required. Or it could separate the amplification into two or more stages, which would be switched in when required and then have lower values of attenuation. (The decision of which way to go depends on other factors, too, such as mass and volume, but we look only at the NF here.) An example of two such cases is shown in Figure 8.4. In the first case, two amplifiers of equal

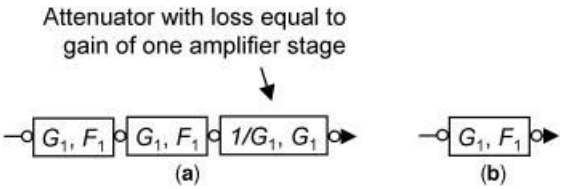


FIGURE 8.4 Same gain attained by means of (a) two-stage amplifier followed by attenuator and (b) single-stage amplifier and no attenuator.

performance are used followed by an attenuator, which undoes the gain of the second amplifier. In the second case, there is just one amplifier and no attenuator. In the first case, the combined NF F is given by

$$F = F_1 + \frac{F_1 - \frac{1}{G_1}}{G_1}$$

while in the second case it is simply

$$F = F_1$$

F in the first case is larger, that is, worse, than in the second case since $G_1 > 1$ and $F_1 > 1$. For the reasonable number of $G_1 = 10$, in the first case $F \approx 1.1F_1$, while in the second case $F = F_1$. So it may be good enough to use the simpler design of the first case.

8.3 HOW TO MAKE AND MAINTAIN PAYLOAD PERFORMANCE BUDGETS

The payload engineer makes and maintains payload-level **budgets** through the entire satellite build (Section 1.1). Budgets are numerical spreadsheets tallying up line items to come up with a bottom line that (hopefully) meets a particular spec. An out-of-spec condition means that the payload engineer has work to do to bring the parameter into spec. The payload engineer must make and keep a budget for every payload parameter to cover every case required in the payload requirements document. The number of cases may run into the hundreds for some budgets for a spot-beam payload.

The budgets reflect the current state of payload design, build, or test and show that the payload will or does meet all its requirements with high probability. Meeting the requirements is the fundamental task of the payload engineer, and the budgets are the tool for seeing how to accomplish this and to show it is accomplished. Properly forming the budgets, especially dealing with the uncertainty inherent in every line item, is a subtle undertaking that we describe in the rest of this section.

8.3.1 Example Budget without Uncertainty: Signal and Noise Levels

We start our discussion of budgets with the construction of one particular budget, a key one. We will go so far as to compute the nominal values of the line items and the bottom line but will delay the crucial but complicated topic of uncertainty to the next subsection.

The **signal- and noise-level budget** tracks the signal level P and the noise level N_0 (one-sided RF noise power spectral density) through the payload. Each line in the budget corresponds to one payload-level element, namely either a payload unit

or a payload-integration element such as waveguide, pad attenuators, and switches. This budget is important because among the payload requirements there will effectively be some requirements on G/T_s and some on EIRP, which are to first order the most important payload parameters. We write "effectively" because the requirements may take slightly different forms from G/T_s and EIRP, respectively; for example, instead of G/T_s it may be receive gain and payload noise figure, and instead of EIRP it may be downlink availability. G/T_s sets the uplink SNR and EIRP the downlink SNR (Section 7.2), as far as the payload can affect these quantities.

Even in the design of active units there is a need to check P/N_0. For example, the noise level entering a component amplifier must not be so high that signal plus noise power drives the amplifier near compression (Section 5.4.2). For another example, the noise level entering an ALC circuit must not be so high that the noise power has a significant effect on where the ALC circuit sets its gain.

In the budget P and N_0 are both computed at the output of every element. Also, the composite gain and NF to that point but referenced to the beginning, is tracked. Additionally but not done in the example below, at any line or lines a bandwidth B could be declared that is of concern or interest there; then the signal-to-noise ratio (SNR) in B at that line, with value $P/(N_0B)$ (see Section 10.6.1), would also be calculated and shown. The budget is normally executed in dB or dB-type units. This particular budget normally has to be done in four cases for every signal path: two cases of signal level into the payload times two cases of preamp mode (ALC and fixed-gain).

The steps in making the budget for the payload are the following:

(a) Define the reference point for the payload's NF calculation. It is the point chosen for the G and T_s calculations. At every line item, the composite NF through to the output of that item will be referenced to the same initial point.

(b) Define T_0 at input to the reference point as the antenna-noise temperature.

(c) Obtain P_0, the power into the reference point, from the payload specification.

(d) Enter into the spreadsheet all the gains and NFs, in order, of all the payload elements in the signal path to be analyzed.

(e) Top the spreadsheet, that is, right above the first element, with the start values of gain of 1 (0 dB) and NF of 1 (0 dB). These stand for the gain and NF through to the output of the reference point and are meaningless except as the necessary starting point of the recursion. Also top the spreadsheet with T_0 and P_0 into the reference point; these are also necessary for the recursion.

(f) Do the following at each succeeding element from first to last. Form the composite gain G through to the output of this element, from the current element's gain and the composite gain of all elements through to the element before the current one. Form the composite NF F from the current element's NF and the composite gain and NF of all elements through to the element before the current one. Compute system noise temperature T_s, always applicable at the

reference point, for the composite of the elements through to this element by applying the formula given below:

$$T_s = T_0 + (F - 1)\,290$$

Compute noise power spectral density N_0 at output of this element by applying the formula given below:

$$N_0 = \kappa G T_s \text{ where } \kappa = \text{ Boltzmann's constant } = 1.379 \times 10^{-20}\ \text{mW/(K-Hz)}$$

Compute P at output of this element, which is as follows:

$$P = GP_0$$

Figure 8.5 gives an example of a signal- and noise-level budget through a simplified signal path of 10 elements (without consideration of uncertainty). The signal power and noise temperature at the input, respectively T_0 and P_0, would be entered in the boxes in the upper left corner. The other items enclosed in thick boxes, near the left side, would be entered into the spreadsheet to describe the individual elements. Note that the source and date of every one of these items are to be entered in the spreadsheet. The middle columns in the table are cumulative.

Variations on this form of the signal- and noise-level budget are sometimes of interest:

- Same form but with $T_0 = 290$, for computing the NF of the payload when the NF is specified.
- Similar form as shown but on just a section of the signal path, as a payload development aid. T_0 is then T_s at output of the part of the signal path that precedes this section.
- Same form as just above but with $T_0 = 290$, for computing the NF of a payload section.

Example of signal- and noise-level budget without uncertainty

T_0 300.00 deg K kappadB −198.60 dB(mW/(K-Hz))
$P_{0:B}$ −30.00 dBm

Element	Elt gain (dB)	Elt NF (dB)	Elt gain (not dB)	Elt NF (not dB)	G (dB)	G (not dB)	F (not dB)	F (dB)	T_s (K)	T_s (dBK)	N_0 (dB(mW/Hz))	P (dBm)	Source of elt gain and NF	Date
Reference point					0.00	1.00	1.00	0.00	300	24.77	−173.93	−30.00		
Element 1	−0.15	0.15	0.97	1.04	−0.15	0.97	1.04	0.15	310	24.92	−173.84	−30.15	Supplier X spec	xx/xx/xx
Element 2	−1.50	1.50	0.71	1.41	−1.65	0.68	1.46	1.65	434	26.38	−173.98	−31.65	Supplier X spec	xx/xx/xx
Element 3	21.00	5.50	125.89	3.55	19.35	86.10	5.19	7.15	1515	31.80	−147.45	−10.65	Supplier Y spec	yy/yy/yy
Element 4	−2.00	2.00	0.63	1.58	17.35	54.33	5.19	7.16	1516	31.81	−149.45	−12.65	Supplier Y spec	yy/yy/yy
Element 5	−3.65	3.65	0.43	2.32	13.70	23.44	5.22	7.18	1524	31.83	−153.08	−16.30	J. Fisher memo	yy/yy/yy
Element 6	−3.65	3.65	0.43	2.32	10.05	10.12	5.28	7.22	1540	31.87	−156.88	−19.95	J. Fisher memo	yy/yy/yy
Element 7	−0.05	0.05	0.99	1.01	10.00	10.00	5.28	7.22	1540	31.88	−158.73	−20.00	Measured in unit test	xx/xx/xx
Element 8	−4.20	4.20	0.38	2.63	5.80	3.80	5.44	7.36	1597	32.01	−160.80	−24.20	Measured in unit test	xx/xx/xx
Element 9	−2.00	2.00	0.63	1.58	3.80	2.40	5.59	7.48	1632	32.13	−162.68	−26.20	Est. based on length	xx/xx/xx
Element 10	−0.15	0.15	0.97	1.04	3.65	2.32	5.81	7.49	1636	32.14	−162.82	−26.35	Supplier X spec	yy/yy/yy

FIGURE 8.5 Example of signal- and noise-level budget without uncertainty.

8.3.2 Dealing with Uncertainty in Budgets

8.3.2.1 Two General Ways of Dealing with Uncertainty A performance budget computes the nominal value of the parameter involved and then degrades the nominal value by a number that reflects uncertainty. The bottom-line result of the budget is then the degraded nominal value, which has a high probability of being achieved over the specified range of on-orbit conditions.

There are two general ways to degrade the nominal value of the parameter value. The most common way is to compute the nominal value's current standard deviation or σ value in the budget; then at the bottom of the budget 2σ or some other multiple of σ is added to the nominal value to create the bottom-line result, but it is subtracted if the difference is a worse case than the sum. This way of obtaining uncertainty is for a budget whose line items all have known or rather well estimated σ. It is discussed in detail in the rest of this subsection. The other way of degrading the nominal value is not to deal with σ values at all but instead to employ a large design margin (Section 8.3.3). This way will not be further discussed.

The actual value of a parameter will be greater than nominal minus 2σ with a probability of 97.7% (Appendix 8.A.5) if the uncertainty has a Gaussian probability distribution. In fact, the uncertainty does not have a Gaussian probability distribution because it consists of the sum of some Gaussian- and some non-Gaussian-distributed component uncertainties. The actual probability distribution is unknowable, since even the assumed distributions of the component uncertainties are only approximations. In summary, the computed combined σ value is a good estimate of the actual combined σ, but the probabilities are only approximations, and the greater the factor applied to σ, the less accurate the approximation. Even though the uncertainty may not be Gaussian, given enough elements in the budget, Gaussian is still a useful approximation (Section 13.2.5.1).

What exactly to use for the nominal value is a nontrivial consideration that is discussed near the end of this subsection on uncertainty.

8.3.2.2 Types of Line-Item Uncertainty In a budget where the σ approach to uncertainty is taken, the σ for the parameter is some combination of σ values of the budget's line items. We must explore what types of line-item uncertainty there are, so we can see how to combine them into the bottom-line uncertainty. To some extent, the types are differentiated by whether the line item itself has a nonzero or zero mean. Line items that have a nonzero mean reflect intentional parts of the design, such as presence of integration elements between units (e.g., waveguide, switches, and hybrids). Some line items that have a zero mean reflect intangibles or unknowables such as unit design immaturity, error in computer model, and manufacturing tolerance. These items have uncertainties that drop out at some point in time. Another group of zero-mean line items apply even on orbit: performance variation over temperature; performance variation from aging and radiation total dose; and error in power measurement, if required. The examples of line-item uncertainty are listed in Table 8.1 with the payload-build stages when they must be taken into account.

TABLE 8.1 Some Types of Budget-Item Uncertainty and When to Take Them into Account

	Payload and Unit Design Stage	Payload Detailed Design and Unit Build Stage	Spacecraft Integration and Test Stage	On Orbit
Error in estimated loss in payload integration elements at room temperature	✓	✓		
Error from unit design immaturity	✓	✓		
Error in computer model	✓	✓		
Error from manufacturing tolerance	✓	✓		
Performance variation over temperature	✓	✓	✓	✓
Performance variation from aging and total radiation dose	✓	✓	✓	✓
Error in power measurement, if required	✓	✓	✓	✓

The good thing is that the various types of uncertainty are mutually uncorrelated (defined in the appendix), so a budget's combined σ is the root-sum-square (rss) of the σs of the various types. As we will see, the hard part is finding how to compute the σ for performance variation over temperature on orbit.

8.3.2.3 Easy Dealing with Some Uncertainty Types The σs for some types of uncertainty are straightforward to deal with, so let us get them out of the way first. Again we reference our discussion to the uncertainty types in Table 8.1, which may not be all types but give the general ideas. Uncertainty in loss in the payload-integration elements at room temperature is of two kinds, one from the uncertainty in waveguide and coax lengths and one from the uncertainty in loss in the other integration elements. The spacecraft engineer who lays out the spacecraft estimates the lengths of waveguide and coax runs, while the payload engineer estimates the error in loss from the other payload-integration elements. Uncertainty from unit design immaturity, computer-model error, and manufacturing tolerance are estimated by the unit engineers and passed on to the payload engineer. Variation over temperature, that is, the temperature-sensitivity coefficients, is provided for the units by the unit engineers and by the payload engineer for the integration elements.

So there are a few σs the payload engineer must estimate himself, from uncertainties provided by the element manufacturers. If the manufacturer just gives an unsigned number, it is best to ask the manufacturer whether it is 1σ, $\pm$ bound with a uniform distribution in between, or something else. If no answer is forthcoming, the number is probably 1σ (and if not, this is a conservative

assumption). If the manufacturer just gives a ± number, it is best to ask whether it is $\pm 1\sigma$, $\pm 2\sigma$, ± bound with a uniform distribution in between, or something else. If the manufacturer writes "worst-case ± *number*" then it is probably $\pm 2\sigma$ (and if not, this is a conservative assumption). The payload engineer then derives the 1σ number from what he is given or his guess. For help on how to do this, see the appendix.

The other fairly straightforward uncertainty type is measurement error, discussed next.

8.3.2.4 Dealing with Error in Power Measurement Whether and how measurement error is taken into account in payload-level power measurements may be driven either by the customer or the manufacturer. There are different ways to take measurement error into account. One way is to require subtraction of a certain amount of measurement error, for example, 2σ, from the measured value. Another way is to ask for the power level above which, with, for example, 84% probability, the actual power will lie. Suppose that the measured power is X and the measurement has an rms error of σ. (Uncertainty in a power meter measurement is internationally stated in terms of an rms value (Agilent, 2009)). The question is then, X minus how many multiples n of σ is the right value to give to the customer. It turns out that n is the same number as if a similar question had been asked about P, as illustrated in Figure 8.6:

$$\Pr(P \geq X - n\sigma) = \text{given probability} \Leftrightarrow \Pr(X \leq P + n\sigma) = \text{given probability}$$

In taking payload measurements used for payload sign-off, the engineer may decide to take more than one measurement if he thinks that the first one yielded a somewhat unrepresentative answer. If say he wants to make three measurements altogether and report the average, the engineer must conscientiously include the very next two measurements after the first one. He may only leave a very bad one out if he knows the probable cause and the customer agrees. Another possible reason for taking several measurements and using the average is to reduce measurement error. If the measurements can be considered to be independent (although

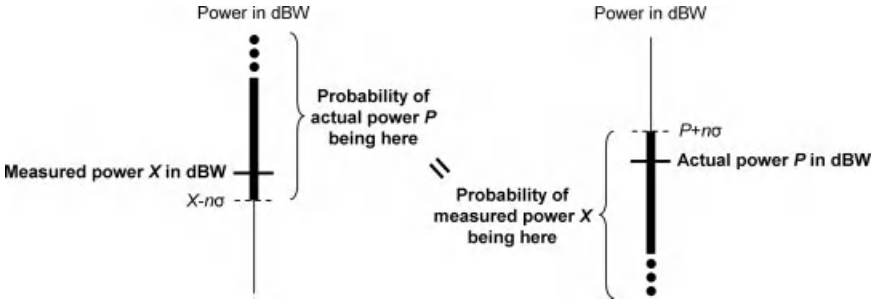


FIGURE 8.6 Probabilities related to output-power measurement.

they should be taken on the same test equipment) then the reduced measurement error is as follows:

Std. dev. of error of n averaged measurements $= 1/\sqrt{n}$ std. dev. of error of one measurement

8.3.2.5 Specifying Environment in Lifetime on Which Payload Performance Must Be Met All right, now for the hard part in dealing with uncertainties in payload performance budgets: computing the σ for the type of uncertainty due to payload performance variation over temperature on orbit over life. Similar in some ways but much simpler to deal with is uncertainty due to aging and radiation. We first consider the spacecraft-level specifications that drive these types of uncertainty and in following subsections translate them into what it means for the payload.

In the spacecraft requirements document, there are two high-level conditions on which the payload performance requirements must be met. The first such condition is the environment: the spacecraft orbit (altitude, inclination, station-keeping, or lack thereof) determines the thermal and radiation environment that the spacecraft will see. The second high-level condition regards the spacecraft lifetime and is variously stated, perhaps dependent on the customer's business model. An example is "at end of life (EOL)," that is, the performance requirements must be met at the stated EOL with x probability. Since (some) units and the thermal subsystem degrade over life (Chapter 2), this means that performance requirements apply at all times through EOL with x probability. Another example of lifetime condition is "average over life," which means, for example, that EIRP must be met $y\%$ of the time averaged from beginning of life (BOL) to EOL. The latter example could even mean that a particular rain model is taken into account, useful for Ku-band and higher frequencies (Section 11.3.6).

8.3.2.6 Dealing with Uncertainty from Aging and Radiation Some units, for example, a TWTA and an oscillator, change characteristics with aging and radiation. Unit engineers have estimates for the probability distributions of these drifts (see examples in that chapter). If the drift from BOL to any later year on orbit may be positive or negative with equal probability then the drift has average or nominal value of zero but an uncertainty that increases with time. If the drift is more likely to be in one direction, it may be compensated (e.g., the TWTA by the CAMP) so that the residual drift is equally likely to be up or down with the years. If such a drift is uncompensated, at whatever year in the life the budget is to be made, the drift's average value for the year will be nonzero (see discussion in Section 8.3.2.9 on what to use for nominal value in such a situation).

8.3.2.7 Converting Thermal Environment in Lifetime into Unit Temperature Variations So now we have only temperature variation to contend with. The manufacturer understands the required spacecraft orbit as implying a whole set of sun direction and sun distance conditions over the year and possibly varying from year to year, for example, when there is no station-keeping. Thermal engineers

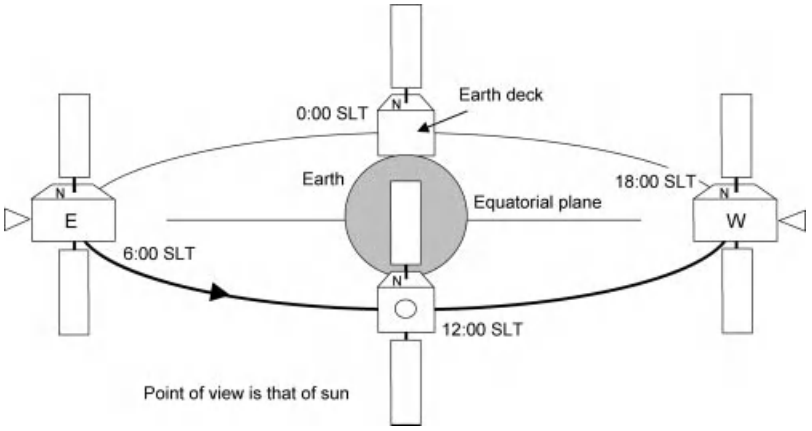


FIGURE 8.7 GEO's diurnal motion at northern hemisphere's summer solstice.

translate the set of sun direction and sun distance conditions into a set of unit temperatures, more specifically, a set of temperatures at unit thermal interfaces (mounting plates) to the spacecraft bus (actually, this is an iterative process between payload and bus as shown in Chapter 1 but we leave that aside). Even if the orbit stays the same over the years, the unit-interface temperatures of units on heat pipes will gradually rise because the thermal subsystem's effectiveness slowly degrades (Section 2.2.1.6).

So the information about unit-interface temperatures is available to the payload engineer to use in budgets. For a particular signal path through the payload, the particular units used are known, so their locations on panels or earth deck and therefore temperatures can be determined. For comparison of two signal paths through the payload, as is done in *C/I* analysis, the task is twice as much work. (Examples of temperature dependence on location are given in Chapter 12.)

Let us see how the thermal environment varies with time so that we can begin to think about how to deal with it. Figure 8.7 shows how a GEO satellite looks to the sun over the course of the summer solstice day in the northern hemisphere (this figure was also given in Chapter 2). The satellite always has its earth deck pointed down to the earth and its thruster pointed away from the earth. Let us suppose that on the earth deck are the LNAs, on the north and south panels are the TWTAs and IMUXes, and on the east and west panels are the OMUXes. The spacecraft drawing at the bottom of the figure represents how the spacecraft looks to the sun at the satellite's local noon: the sun sees the thrusters and the north panel at a sharp angle (23.44°). So the TWTAs and IMUXes on the north panel get a little sun—actually, they get the same amount of sun at all times that day. The spacecraft drawing on the right side of the figure is 6 h later, and now the sun is almost perpendicular to the west panel, actually at 66.56° to the panel, so the OMUXes on the west panel get a lot of sun. Another 6 h later, at midnight, and the sun is at 66.56° to the earth deck, so the LNAs get a lot of sun. Another 6 h later, and the sun is at 66.56° to the east panel, so the OMUXes there get a lot of sun. So even just over a day, the units on the earth deck and the east and west panels have a wide variation in illumination.

At other times of the year, the situation is similar. At the winter solstice, the south panel instead of the north panel gets a little sun all day; the earth deck and east and west panels have the same situation as in summer. At the equinoxes, the north and south panels get no sun at all and the earth deck and east and west panels get direct sun at times; both equinoxes are the same in this regard.

As the years go by, a unit's daily temperature swing on any particular day of the year will have an increasingly high average temperature (Chapter 12).

8.3.2.8 Dealing with Performance Variation with On-Orbit Temperature Therefore, the absolute best way to incorporate the variation in payload performance with temperature into the sensitive budgets is to make a version of the budget at every minute of the spacecraft life. If the payload requirements document says the performance requirements have to apply at EOL, only the budgets over the last year would be needed.

This way is of course totally infeasible. So we come to the second-to-the last condition on making budgets, and that is that the chosen set of budgets must be computationally feasible.

Now, it is almost as bad to have conservative budgets as optimistic ones. Having optimistic budgets is clearly bad—they can lead to payload design mistakes that have to be corrected at a late stage or to payload performance that does not meet customer requirements. However, having conservative budgets can mean that spacecraft capacity is being wasted that could have been used to offer higher performance or additional capabilities to the customer, or it could mean that the extra mass drives the launch vehicle to be unnecessarily large, or it could mean that the extra resources could never have been implemented to begin with and the manufacturer could have saved money. So the last condition on making budgets is that they must be conservative if they cannot be accurate, but as little conservative as possible.

So some simplifications must be made. There are some budgets for which a simplification is easy to formulate. Some payload parameters have variations whose worst case seems readily accommodated by the unit engineers. G/T_s is a somewhat similar parameter except that it spans many units. G of a passive antenna will not change over life. Since T_s is set by relatively low-power devices, it is often not a tough design decision to add the relatively few "extra" resources to accommodate its lifetime temperature variation. For these parameters, there is no need to use a σ value; in the one version of the budget, the line items are entered at their worst temperature condition. Another case with easy simplification is when the EOL temperature variation σ has negligible effect on the rss of all the σs in the budget.

The difficult budget simplifications are for EIRP. Payload performance is generally worse at hot than at cold. Amplifiers perform worse at hot. Some units or other payload elements may need to be temperature-compensated to maintain a nearly constant performance over temperature. A HPA changes characteristics with temperature (also with aging and radiation) (Chapter 6). If the payload transmits noninterfering beams, it may be sufficient to just "throw HPA power" at the problem. However, if bus power is tight or the payload transmits overlapping spot beams, some means of compensation by the preamp must be implemented. In summary,

with temperature and the years, the payload changes performance and possibly also the way it functions.

So for the EIRP budget, to make good simplifications, we have to consider more closely the thermal environment variation over the day and year, with the help of Figure 8.7 again. One feasible simplification is to choose the worst day of the year in terms of payload performance and let it stand in for the whole year. What day this is depends on the spacecraft layout and the payload technology. What is left then are the diurnal variation and the lifetime requirement (i.e., whether the payload specs apply at EOL, at a few years earlier, or to some average or percentage of the time over the years).

Let us assume that in the signal path under consideration, only one unit plus associated integration waveguide is on the earth deck or east or west panel; such is the case in the examples in Chapter 12. For simplicity's sake, let us look now just at EOL. The temperature variation of that unit and waveguide over the worst day will have a **probability density function (pdf)**, defined in Appendix 8.A.2) like that given in Appendix 8.A.6, where for half the day the temperature is at its daily minimum and for the other half of the day the pdf increases with temperature to the day's maximum. The mean unit temperature is about 0.32 of the way from minimum to maximum, and the temperature's standard deviation σ is about 0.39 of the maximum minus the minimum. The nominal budget should be computed with the temperature-varying unit and waveguide at their mean temperature over the day. The budget would be computed again at 0.71 of the way from minimum to maximum temperature, and the difference between the budgets would be the σ due to temperature variation. The second budget would not be further used; only the σ would be entered into the budget at the mean temperature. All the other units and integration elements will have constant temperature.

8.3.2.9 Nominal Value We have defined variation as being about a nominal value, and exactly what the nominal value should be is a nontrivial consideration.

When the lifetime requirement is "at EOL," clearly the nominal value should be for the EOL. Most of the types of uncertainty in performance of a unit or payload-integration element have the same value whether the performance is being predicted for EOL or BOL. For those types, the nominal performance value should be the expected value or current best estimate, not a conservative estimate, since the conservatism is taken care of in the addition or subtraction of 2σ in the budget's bottom line. The nominal antenna gain in a particular direction is the gain averaged over antenna-pointing error. There are two types of performance uncertainty that sometimes do have a different estimate for EOL from that for BOL: that due to radiation and aging and that due to unit temperature rising over life from the thermal subsystem aging. In our payload example in Chapter 12, radiation and aging cause a different expectation with the years but temperature rising causes only a negligible difference. Most TWTAs are expected, based on their test data, to put out a little bit less power at EOL than at BOL but some are expected to put out a little more. Normally, the lifetime requirement of "at EOL" is meant to be a conservative requirement. For a TWTA whose performance is expected to improve, the BOL performance should perhaps be used in the budgets.

When the lifetime requirement is "average over life," the nominal value should be the expected performance in the middle of life, assuming that drifts are linear over life. A suggestion of how to handle this in the bottom line of a budget, where normally 2σ is either subtracted or added, is to use $|\delta| + 2\sigma$ or $|\delta| - 2\sigma$ instead, where δ is the drift expected to occur from the middle of life to EOL. When the payload specification has a link availability requirement instead of EIRP requirements and the frequency band is 10 GHz or higher, $|\delta|$ up to about 0.1 dB can be ignored because availability at EOL and at BOL will effectively average out to availability in the middle of life.

8.3.2.10 Combining Line-Item Uncertainties Recall that the various types of uncertainty are mutually uncorrelated (Section 8.3.2.2), so a budget's combined σ is the rss of the σs of the various types. Since the budget items are in dB, it is convenient if also the uncertainties are in dB. Some uncertainties are clearly available in dB, for example, waveguide loss per foot and coupler loss. If, however, an uncertainty is not in dB it can easily be turned into dB. If an uncertainty is given as a power ratio X, $+u$, $-v$, turn this into the fractions $+u/X$, $-v/X$. Then let ε equal either one of these fractions. The uncertainty represented by the fraction is approximately equal to $(10/\ln 10)\varepsilon$ dB or, more accurately, $(10/\ln 10)(\varepsilon - 0.5\varepsilon^2)$ dB.

Figure 8.8 shows an example of a simplified signal-level budget carried through three units and their integration hardware. Normally the noise level would also be

Unit 1 and associated integration elements

				Low signal		High signal			
Element	Elt gain (dB)	Cum G (dB)	σ (dB)	Nominal P (dBm)	Low P (dBm)	Nominal P (dBm)	High P (dBm)	Source	Date
Reference point		0.00	**0.00**	**−60.00**		**−30.00**			
Connector	−0.15	−0.15	0.05	−60.15		−30.15		Supplier X spec	xx/xx/xx
LNA	10.00	9.85	0.00	−50.15		−20.15		K. Timm memo	tt/tt/tt
Design immaturity	0.00	9.85	0.40	−50.15		−20.15		K. Timm memo	tt/tt/tt
Measurement error	0.00	9.85	0.10	−50.15		−20.15		K. Timm memo	tt/tt/tt
Cum rss of σ's			**0.42**						
Signal level				**−50.15**	**−50.98**	**−20.15**	**−19.32**		

Unit 2 and associated integration elements

				Low signal		High signal			
Element	Elt gain (dB)	Cum G (dB)	σ (dB)	Nominal P (dBm)	Low P (dBm)	Nominal P (dBm)	High P (dBm)	Source	Date
Reference point		0.00	**0.42**	**−50.15**		**−20.15**			
Connector	−0.15	−0.15	0.05	−60.15		−20.30		Supplier X spec	xx/xx/xx
Waveguide	−1.50	−1.65	0.20	−61.65		−21.80		J. Frank memo	yy/yy/yy
HPA	30.00	28.35	0.40	−31.65		8.20		Supplier Z spec	zz/zz/zz
Aging and radiation	0.00	28.35	0.25	−31.65		8.20		Supplier Z spec	zz/zz/zz
Measurement error	0.00	28.35	0.20	−31.65		8.20		Supplier Z spec	zz/zz/zz
Cum rssof σ's			**0.69**						
Signal level				**−31.65**	**−33.03**	**8.20**	**9.58**		

Unit 3 and associated integration elements

				Low signal		High signal			
Element	Elt gain (dB)	Cum G (dB)	σ (dB)	Nominal P (dBm)	Low P (dBm)	Nominal P (dBm)	High P (dBm)	Source	Date
Reference point		0.00	**0.69**	**−31.65**		**8.20**			
Waveguide	−1.20	−1.20	0.05	−32.85		7.00		J. Frank memo	yy/yy/yy
OMUX	−0.50	−1.70	0.05	−33.35		6.50		C. Hendrix PDR	zz/zz/zz
Over temp in life	0.00	−1.70	0.07	−33.35		6.50		C. Hendrix PDR	zz/zz/zz
Measurement error	0.00	−1.70	0.10	−33.35		6.50		C. Hendrix PDR	zz/zz/zz
Cum rssof σ's			**0.71**						
Signal level				**−33.35**	**−34.76**	**6.50**	**7.91**		

FIGURE 8.8 Example of simplified signal-level budget (with uncertainty).

carried in such a budget but for the sake of clarity we leave this out. Both high and low signal levels are carried. A nominal signal level is carried, and the low case is decreased further by the 2σ of uncertainties and the high case is augmented by the addition of this amount. The nominal levels at the output of unit 1 form the "reference point" values at input to unit 2, and so on. The combined 2σ increases as we go through the units. All the uncertainties are uncorrelated, so they are rss'ed together. Unit 3 is on the earth deck or east or west panel. The values entered in the left section of the table are not meant to be representative.

8.3.3 Keeping Margin in Budgets

Performance budgets usually carry one or more kinds of **margin**. Margin is headroom. The various kinds of margin are carried for different purposes. Most budgets show only one or two, depending on the purpose of the budget.

- ***Risk Margin***: This margin is carried in certain budgets throughout the program by the satellite manufacturer. It is a measure of the uncertainty that the engineer feels in his calculations at the point in time. It is usually a percentage of the total budget. Based on his experience on other programs, the engineer knows about how much his current numbers could grow from that time on. (Numbers can also decrease, but when a decrease is benign, as is almost always the case, the engineer does not need to protect himself.) The risk margin decreases over the course of the program; when the unit or payload is completed and measured, this margin is zero dB.
- ***Design Margin in High-Error Budgets***: This kind of margin is only used in the few budgets in which the individual terms cannot be well estimated or measured. An example is PIMs.
- ***Customer-Required Margin***: This kind of margin is not commonly required. If required, it would be for a clear and definite purpose for the customer, since requiring the payload to carry "extra" margin is an expensive proposition.
- ***Bottom-Line Margin***: This is the excess margin after taking into account uncertainties in the budget items and any of the above other margins. The payload engineer must ensure that this is non-negative at all times.

Choosing the margins other than the bottom-line margin is a delicate balancing act. If they are too small, in the end the payload may not meet its specifications. On the other hand, it is almost as bad to carry too much margin. The margins must be chosen with a reasonable amount of risk in mind.

The manner in which risk margins are held at both payload and unit levels should be consciously decided on. The payload engineer must either know how the unit engineers are doing it or tell them how they should do it; in either case, he must depend on their openness in the matter. It is wasteful if the unit engineers carry their own risk margin and the payload engineer carries an additional margin for those units for the same purpose.

8.3.4 Maintaining Budget Integrity

The payload budgets must be kept current and accurate, to reflect the current state of knowledge of the payload. The reason is that many decisions must be made throughout the program on the basis of the budgets: decisions on unit specifications and acceptance, on spacecraft layout, on what kind and how much waveguide to buy, and so on.

The payload budgets need not be updated every day, but ideally every time that better numbers are available for entries. First numbers are the result of the proposal analysis; new numbers become available at detailed payload design, detailed unit design, unit test, integration, final test. The budgets must at least all be updated for PDR, CDR, and final acceptance review.

The individual elements' gain and noise figures would be entered into the spreadsheet. The source and date of every one of these items are to be entered in the spreadsheet. Practical experience shows that this is absolutely essential at payload level to avoid a lot of head-scratching and repeated work. There are so many payload performance budgets, there are so many entries in each budget, and the entries must be kept up to date as the satellite program progresses that without notes it is impossible to remember whether or not individual elements are due for an update or not.

8.4 HIGH-POWER AMPLIFIER TOPICS

8.4.1 How to Know If HPA Nonlinearity Should Be Specified on C/3IM or NPR

It is useful to specify the HPA nonlinearity on the basis of either its C/3IM or NPR characterization (both defined in Section 6.2.2) when the operational signals into the HPA will be much like the characterization test signals, but in other circumstances these characterizations are not useful. This was stated in Section 6.2.2 but the explanation was deferred to here since it rests on probability theory, first presented in this chapter's appendix.

The C/3IM test signals are two equal-power sinewaves of different frequencies, and the NPR test signals are a large number of equal-power sinewaves equally spaced in frequency. The NPR test signals are an approximation to white Gaussian noise that has been brickwall-filtered (Section 10.3.4.2). The pdf of the instantaneous total power of the combined test signals is shown in Figure 8.9 for both characterizations. Both plots in the figure are scaled so that the long-term average total power is 1. For C/3IM, the pdf curve is fairly flat over most of the interval [0,2] but rises to infinity at 0 and 2 and is zero beyond 2. For NPR, the pdf curve falls exponentially over the interval from 0 to a large number. These are quite different.

The plots in Figure 8.9 tell us how the HPA will be exercised in test. Recall that the HPA operating point has as its P_{in} the long-term average input power; thus, the operating point corresponds to the x-value of 1 in the plots of Figure 8.9. The HPA will be tested on a wide range of operating points. For C/3IM, the HPA will be

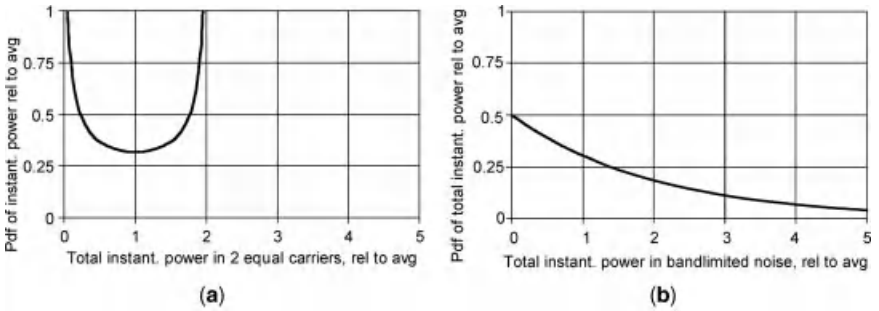


FIGURE 8.9 Probability density functions of instantaneous total power in (a) two equal-power carriers and (b) bandlimited Gaussian noise.

driven only up to 3 dB beyond each operating point. For NPR, the HPA will be driven as far beyond the operating point as is feasible for the HPA, indeed at least 3 dB above the operating point about one third of the time. So C/3IM and NPR characterize the HPA at any given operating point in very different ways. Care must be taken in deciding if the operational signals into the HPA will be sufficiently close to either set of test signals for C/3IM or NPR to justify choosing one of these characterizations for the HPA nonlinearity.

Both C/3IM and NPR are bad characterizations when the operational input to the HPA will be one strong signal. The pdf of the instantaneous power in one carrier is shown for QPSK modulation in Figure 8.10. Two cases of the RRC pulse rolloff α (Section 10.3.4.3) are shown, namely infinity and 0.2. For both, the pdf is heavily weighted around 1, the long-term average power, which is clearly a very different situation from either curve in Figure 8.9.

8.4.2 What HPA Nonlinearity Does to Signal

The intermodulation products (IMPs) of the HPA are mostly inband or near-out-of-band. Since the inband ones cannot be removed, they must be low enough not to be too big of a problem. **Near-out-of-band** means outside the channel but in nearby channels.

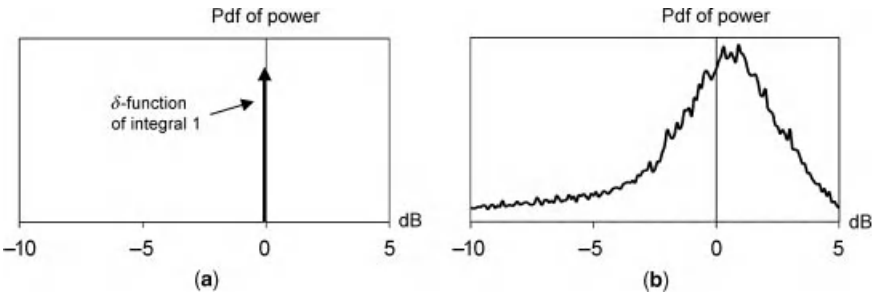


FIGURE 8.10 Pdf of power of QPSK signal with (a) rectangular pulse and (b) RRC pulse with $\alpha = 0.2$.

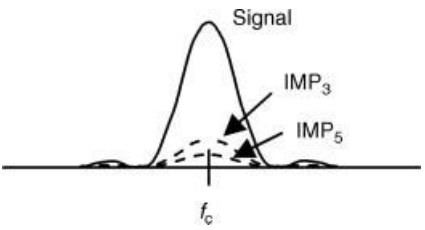


FIGURE 8.11 Example of signal and its 3rd- and 5th-order IMPs.

There are different ways to think about what the nonlinearity does to the signal. The most direct and informative way is to think about the IMPs. Then it is clear why spectrum-spreading and power robbing occur and what they are.

8.4.2.1 HPA in Terms of Intermodulation Products

Case 1: HPA Amplifies One Signal If the HPA amplifies only one signal, the HPA outputs the main signal plus odd-number IMPs, for all positive real numbers, at the carrier frequency f_c. Usually, the 3rd- and 5th-order IMPs are the only ones big enough to matter. The IMPs become less powerful with increasing order. The shape of the IMPs' spectra depends on the modulation pulse (e.g., RRC with $\alpha = 0.2$). For the simple-to-compute case of the rectangular pulse, the 3rd- and 5th-order IMPs have the same shape spectrum as the original signal, as illustrated in Figure 8.11.

Case 2: HPA Amplifies Two Signals of Equal Power and Same Spectrum The case of two signals is more complicated. Again, it is the 3rd-order IMPs and some of the IMPs of each odd higher order that overlap the signals. The signals and their 3rd-order IMPs are shown in Figure 8.12. There are four

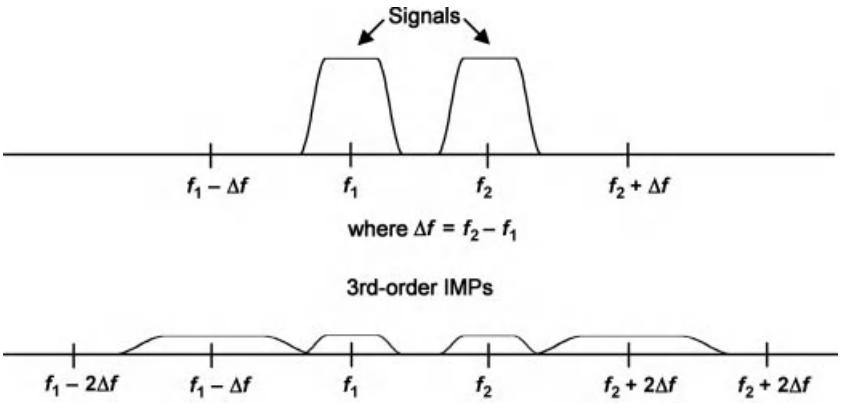


FIGURE 8.12 Example of two equal signals and their inband and near-inband 3rd-order IMPs.

3rd-order IMPs. A 3rd-order IMP overlays each signal and is almost the same as that signal but of opposite sign; it is effectively a power reduction in the signal, so it is not separately detectable. The two other 3rd-order IMPs are offset, and each has twice the power as one of the other kind. The spectrum of each is nearly the same as that of the other kind of IMP but twice as wide (latter assuming that the signals were modulated with independent data clocks).

Case 3: HPA Amplifies Two Unequal Signals When two signals are unequal in power and are amplified together, the HPA will suppress the smaller signal relative to the larger one. An ideal hard limiter suppresses a much smaller signal by 6 dB. The HPA is not a hard limiter, but in general it cannot be operated at saturation when there are two signals. The input backoff typically must be at least 3 dB, where still the suppression must be taken into account in calculations of output-power specification compliance for the weaker signal. The "missing" power of the weaker signal goes into the IMPs.

Figure 8.13 shows two unequal signals and their 3rd-order IMPs. The IMPs now have unequal power. The power ratio of the two IMPs that overlap the output signals is that of the output signals, but, as before, they are not noticeable. The IMPs on the sides are intermediate in size, with the one next to the larger signal being larger than the one next to the smaller signal.

Case 4: HPA Amplifies a Few Signals When there are more than two signals, say three or four, it is best if they are unequally spaced across the band, to minimize the number of 3rd- and 5th-order IMPs overlapping the signals. Reference to Figure 8.13 and some thought show the danger of spacing the signals equally, especially if they are of different power levels: if three signals are evenly spaced and one is much smaller than the other two, a 3rd-order IMP overlapping the smallest signal could have size on the same order as the smallest signal. If the signals have to be evenly spaced or randomly spaced, it is best if the ratio of powers of the various

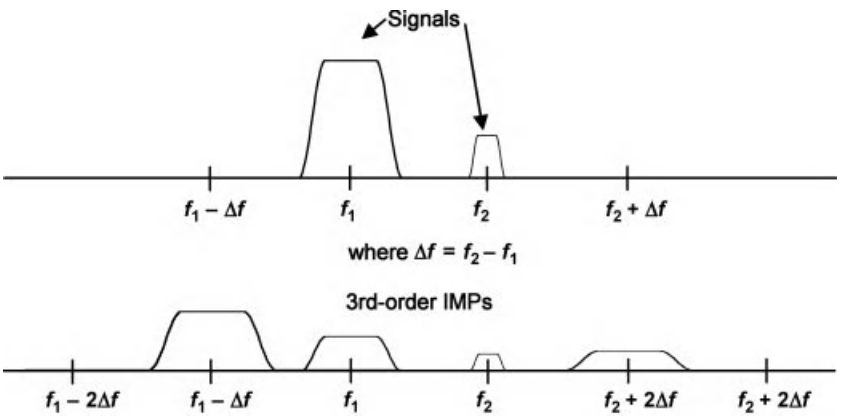


FIGURE 8.13 Example of two unequal signals and their 3rd-order IMPs.

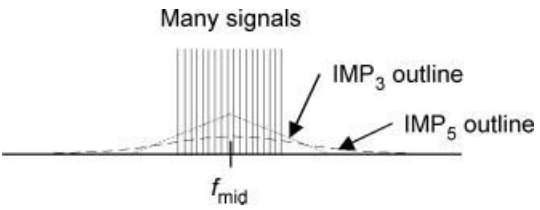


FIGURE 8.14 Many signals and outlines of their 3rd- and 5th-order IMPs.

signals to each other be within a limited range, even if some signals are thereby overpowered. Operating the HPA well backed off is an alternate solution. The HPA that will always be run well backed off is optimized by the HPA manufacturer differently from the HPA that will always be run at saturation or just a few dB backed off.

Case 5: HPA Amplifies Many Evenly Spaced, Independent Signals of Equal Power and Same Spectrum Another simple case is a very large number of evenly spaced, independent signals of the same power and bandwidth. Figure 8.14 shows the spectra of the intended signals and the outlines of the spectra of their 3rd- and 5th-order IMPs. Since the signals are independent, the overall shape of the kth IMP of the group of signals is that of the original group convolved with itself $k-1$ times.

In practice, the HPA must be well backed off in this case. The HPA that will always be run well backed off is optimized differently by the HPA manufacturer from the HPA that will always be run at saturation or just a few dB backed off.

Case 6: HPA Amplifies Many Unequal Signals What the 3rd- and 5th-order IMPs look like for any other signal-set case can roughly be evaluated from the other cases presented here but a good evaluation requires simulation.

8.4.2.2 HPA in Terms of Spectrum-Spreading We saw in case 1 that the HPA creates odd-order IMPs that overlay the input signal's spectrum and are wider. Thus, the output signal's spectrum is altogether wider than that of the input signal. The HPA has caused the signal's spectrum to "spread." An example is shown in Figure 8.15 for the RRC pulse with $\alpha = 0.35$ (Section 10.3.4.3). The loss of power in the intended signal is visible, and this power has gone into IMPs.

In the case where there was an original signal spectrum that was narrowed by pre-HPA filtering, the HPA would cause some of the outer edges of the signal spectrum to come back, at a lower level. This is called "spectral regrowth."

The spreading of the spectrum is a concern for the adjacent channels (adjacent-channel interference or ACI) so must be controlled.

8.4.2.3 HPA in Terms of Power Robbing **Power robbing** is a term that reflects the fact that the HPA suppresses a weaker signal relative to a stronger signal when two unequal-power signals are input to the HPA. This occurs when the

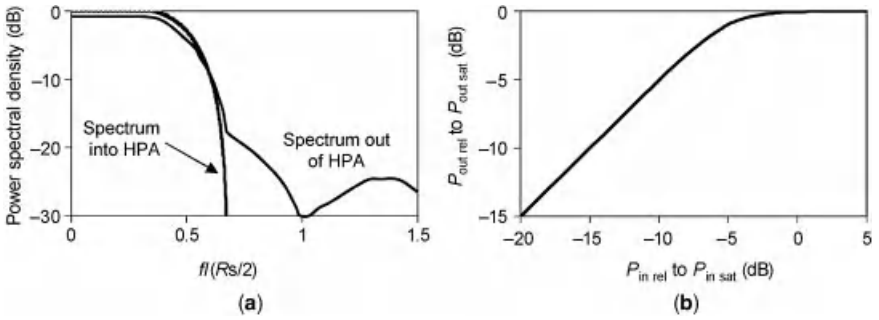


FIGURE 8.15 Spectral regrowth: (a) one side of signal spectrum, into and out of HPA, and (b) assumed P_{out} versus P_{in} of HPA.

HPA operates in its nonlinear-amplification regime. The closer to saturation the HPA is, the more the suppression is. It appears that the stronger signal is robbing power from the weaker.

It is not usually an issue when there is only one signal plus noise, but it can be when the signal arriving at the satellite is on the low end of the allowed range. Then the signal-to-noise ratio is at its worst. If the noise power in the passband of the HPA is larger than that of the signal, at HPA output the signal will be suppressed relative to the noise.

To get a good idea of the power robbing in a particular case, it is necessary to perform a simulation or measurement (Chapter 13). However, the following will give a very rough idea of what to expect (Jones, 1963). It is based on two signals and noise going into an ideal hard limiter (HL).

- When the noise is much stronger than the signal(s), the HL degrades the SNR(s) of the signal(s) by 1 dB.
- When one signal is much stronger than both the other signal and the noise, the HPA increases the SNR of the strong signal by 3 dB and decreases the SNR of the weak signal by 3 dB.

In the second case, the SNR of the stronger signal actually gets enhanced. But for PSK modulations this creates no advantage, since all that has happened is that the radial component of the noise is suppressed, but this component has nothing to do with signal detection (Section 10.5.4).

8.4.3 How to Ease Payload Integration of Combined TWTAs

In a transponder where the signal is split, fed to two or more TWTAs, and combined in phase and where the post-TWTA RF line is waveguide, the lengths of the waveguides to the combiner must be closely matched. If the lengths are not matched, the path delays are different. The TWTAs themselves can be made phase-matched to within a few degrees, even though the absolute phase shift through a TWTA is

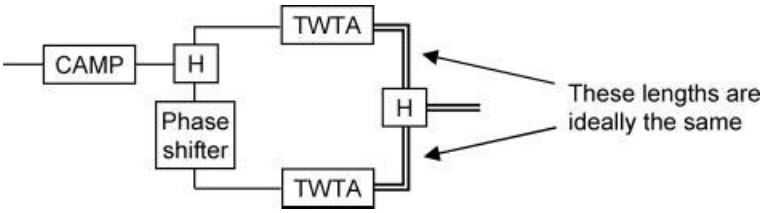


FIGURE 8.16 In-phase combining of TWTA outputs.

thousands of degrees. The situation is illustrated in Figure 8.16. If the channel band is wide (on the order of 3% of the carrier frequency) then the phase shift at the upper end of the channel band relative to the phase shift at the lower end of the channel band may be many degrees different between the two waveguide runs. This would cause serious degradation in the combined signal. A good solution is to adjust the relative length of the two pieces of pre-TWTA coaxial cable to provide a compensating differential delay in the waveguide.

The technique is as follows. We first figure out how many degrees of phase shift the lower and upper channel frequencies will see in the extra length of waveguide, and we take the difference. By "extra length" we mean the length of the longer piece of waveguide minus the length of the shorter piece. Then we find the length of coax that will provide the same difference between the frequencies. (The absolute phase shift at each of the two end frequencies will be different in coax from what they are in the waveguide.) Between the two end frequencies the match will not be perfect but almost.

In detail the calculation is as follows. In rectangular waveguide excited by the usual TE_{10} mode, the wavelength λ_g at propagating frequency f above the waveguide's cutoff frequency f_c is as follows (Ramo et al., 1984):

$$\lambda_g = \frac{c/f}{\sqrt{1-(f_c/f)^2}} \text{ where } c \text{ is the speed of light in free space}$$

If d is the extra length of waveguide then the phase shift at the channel band's upper frequency f_2 minus the phase shift at the lower frequency f_1 is as follows:

$$\Delta\varphi_{wg} = \left(\frac{d}{\lambda_{g2}} - \frac{d}{\lambda_{g1}}\right) 360 \text{ in deg where } \lambda_{gi} = \text{guided wavelength at frequency } f_i,\ i = 1, 2.$$

Now, in coax the guide wavelength is simply proportional to the free-space wavelength. In coax with dielectric constant ε_r excited by the usual TEM mode, the guide wavelength at frequency f is as follows:

$$\lambda_g = \frac{c/\sqrt{\varepsilon_r}}{f} \text{ or } \frac{1}{\lambda_g} = \frac{f}{c/\sqrt{\varepsilon_r}}$$

In a piece of coax of unknown length m, the phase shift at f_2 minus the phase shift at f_1 is as follows:

$$\Delta\varphi_{\text{coax}} = m\left(\frac{1}{\lambda_{\text{g2}}} - \frac{1}{\lambda_{\text{g1}}}\right)360 \text{ in deg where now } \lambda_{gi} \text{ applies to the coax}$$

So to make $\Delta\varphi_{\text{coax}} = \Delta\varphi_{\text{wg}}$ we simply use the length m of coax given by

$$m = \frac{\Delta\varphi_{\text{wg}}/360}{\left(\dfrac{f_2 - f_1}{c/\sqrt{\varepsilon_{\text{r}}}}\right)}$$

For example, a channel band is the frequency band between 12.2 and 12.7 GHz. WR75 waveguide has the cutoff frequency 7.868 GHz and a 141 coax has a dielectric constant of 2.04. An 8-cm extra length of waveguide can be compensated by a 7.23-cm extra length of coax. The phase shift over the channel band, relative to the phase shift at 12.2 GHz, is shown in Figure 8.17 for both the waveguide and the coax. Both the waveguide and the coax see a phase shift of about 62° across the band. The difference across the band of the relative phase-shift plots is shown in Figure 8.18. The maximum difference across the band, that is, the error in the compensation, is only about 0.21°.

A 360°-capable phase shifter is required in one of the arms of the combiner for several reasons: (1) the waveguide and coax pieces can never be cut exactly right; (2) the phase shift in waveguide and coax bends cannot be accurately predicted, to the author's knowledge; and (3) the guide wavelengths will shift a little when the waveguide goes from air to vacuum (Section 4.3.2.3). If the use of a 360°-capable phase shifter is not practical then two of the 180°-capable ones can be used.

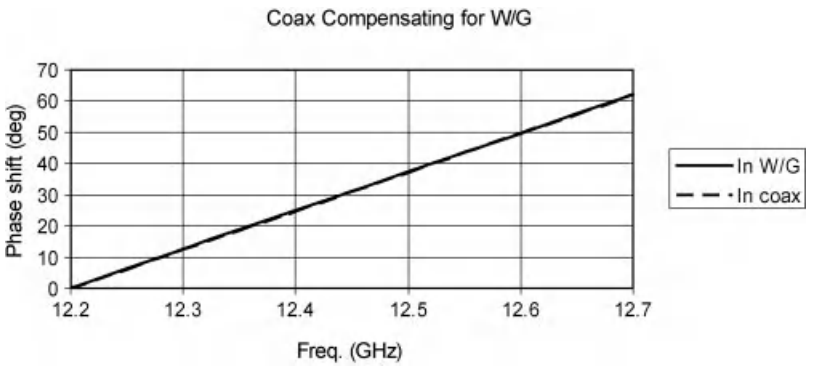


FIGURE 8.17 Relative phase shift over the band in a length of waveguide and in compensating coax length.

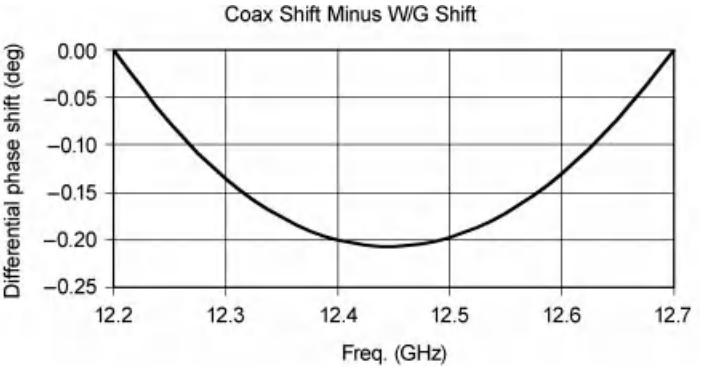


FIGURE 8.18 Compensation error over the band.

8.5 HOW TO AVOID MONTE CARLO SIMULATIONS ON GAUSSIAN RANDOM VARIABLES

Gauss–Hermite integration is a tool that can replace most Monte Carlo simulations over independent Gaussian random variables with a quick calculation. What is being evaluated in either method is a function whose independent variables have independent Gaussian probability distributions. Gauss–Hermite integration is a weighted sum of the function evaluated at a few points, so the computational burden is much less. This is especially a benefit when the function is being evaluated over many random variables including the Gaussian ones or when the function is evaluated as just part of a longer calculation. Following are some examples of payload calculations where Gauss–Hermite integration is a possibility:

- EIRP as a function of antenna pointing error
- EIRP as a function of repeater-caused uncertainty
- Loss in combining HPA outputs aligned in phase, as a function of phase misalignment
- HPA output power or phase shift, averaged over the input noise or over the other signals into the HPA which together look like noise (Section 13.2.4).

The way to perform a Gauss–Hermite integration is to first perform it on say 6 points and then on 10 points and to check if the difference in the answers is within a chosen error bound. If so, the 6 points were enough. If not, the procedure must be repeated with say 10 and 16 points. It has been the author's experience that 6 points is enough for the functions she has used this on, which included the error function and link availability due to rain at Ku-band. If the function to be evaluated is dependent on other variables besides the Gaussian ones, try 6 and 10 points on sets of the other variables which represent the extreme cases of the function; if 6 points is enough here then it will be enough elsewhere.

TABLE 8.2 Gauss–Hermite Integration Abscissas and Weights

$\pm x_i$	w_i	$\pm x_i$	w_i	$\pm x_i$	w_i
$n=6$		$n=10$		$n=16$	
0.6167065900	4.088284695e-1	0.4849357074	3.446423349e-1	0.3867606043	2.865685212e-1
1.889175877	8.861574602e-2	1.465989094	1.354837030e-1	1.163829100	1.583383727e-1
3.324257433	2.555784402e-3	2.484325841	1.911158050e-2	1.951980345	4.728475235e-2
$n=8$		3.581823482	7.580709344e-4	2.760245047	7.266937603e-3
0.5390798112	3.730122577e-1	4.859462827	4.310652630e-6	3.600873623	5.259849265e-4
1.636519041	1.172399076e-1			4.492955301	1.530003216e-5
2.802485861	9.635220121e-3			5.472225705	1.309473216e-7
4.144547185	1.126145384e-4			6.630878196	1.497814723e-10

Source: After Abramowitz and Stegun (1965).

The formula for Gauss–Hermite integration is as follows:

$$\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\infty} e^{-y^2/2\sigma^2} f(y)dy \doteq \sum_{i=1}^{n} w_i f(\sigma X_i)$$

where the abscissas x_i and the weights w_i for a few n are given in Table 8.2 (derived from Abramowitz and Stegun (1965)). See Abramowitz and Stegun (1965) for more sets of abscissas and weights for n up through 20.

APPENDIX 8.A

8.A.1 Elements of Probability Theory for Payload Analysis

After Fourier analysis, probability theory is the field of mathematics most commonly applied in specification, development, and sell-off of the communications payload (and indeed the end-to-end communications system). We present definitions and properties, then practical and positively useful tips on applying probability theory. Further topics can be found in the textbook (Papoulis, 1984) or any of its later versions or in just about any textbook on elementary probability theory.

8.A.2 Definition of Random Variable and Probability Density Function

We summarize here the definitions of the probability terms most commonly needed in communications analysis involving the payload (or indeed the end-to-end communications system).

A "probability space" is made up of two things: a "space" or set $\mathcal{S}$ of elements and a "probability function" Pr defined on "events" or subsets of $\mathcal{S}$ that has the following properties (Papoulis, 1984):

(1) $\Pr(\mathcal{A}) \geq 0$ where $\mathcal{A}$ is a subset of $\mathcal{S}$
(2) $\Pr(\mathcal{S}) = 1$
(3) $\Pr(\mathcal{A} \cup \mathcal{B}) = \Pr(\mathcal{A}) + \Pr(\mathcal{B})$ if $\mathcal{A} \cap \mathcal{B} = \emptyset$, where $\mathcal{A}$ and $\mathcal{B}$ are subsets of $\mathcal{S}$, " $\cup$ " means "union," " $\cap$ " means "intersection," and $\emptyset$ is the null set.

The elements of $\mathcal{S}$ are also events, the "elementary events." $\mathcal{S}$ could, for example, be the interval [0,1] of real numbers between and including 0 and 1.

Roughly speaking, a **random variable** or **random number** X is a real-valued function defined on the elements of a probability space $\mathcal{S}$ (see Papoulis (1984) for a complete definition). Then on a subset of $\mathcal{S}$, X takes on a set of values. We can invert this and define a subset of $\mathcal{S}$ by the set of values that X takes on it:

$$\Pr(X = x) \triangleq \Pr(\text{event on which } X = x)$$

where x is a value of X and "$\triangleq$" means "is defined as." Similarly,

$$\Pr(X \leq x) \triangleq \Pr(\text{event on which } X \leq x)$$

The "(cumulative) distribution function" P_X of the random variable X is defined by

$$P_X(x) \triangleq \Pr(X \leq x) \text{ for all values } x \text{ of } X$$

When there is no ambiguity, the subscript X is usually omitted. If the distribution function P_X is continuous, then X is said to be of "continuous type," and if the function is a staircase-type, X is said to be of "discrete type" (Papoulis, 1984). The "mixed type" comes up now and then. The above definition of distribution function applies to both continuous and discrete random variables. The **(probability) density function** (**pdf**) p_X for a continuous random variable is defined by

$$p_X(x) \triangleq \frac{d}{dx}\Pr(X \leq x) \text{ for all values } x \text{ of } X$$

For a discrete random variable it is defined by

$$p_X(x_i) \triangleq \Pr(X = x_i) \text{ for all values } x_i \text{ of } X$$

Sometimes the subscript X is omitted.

8.A.3 Mean, Standard Deviation, and Correlation

The properties of a random variable X most often used in payload analysis are the **mean** m_X and **standard deviation** δ_X:

$$\begin{aligned} m_X &= \text{mean of } X \triangleq \int_{\mathcal{S}} x p_X(x) dx \\ \sigma_X &= \text{standard deviation of } X = \sqrt{\text{variance of } X} \\ &\quad \text{where variance of } X \triangleq \int_{\mathcal{S}} (x - m_X)^2 p_X(x) dx \end{aligned}$$

The subscript X may be dropped when the meaning is clear.

Often we are interested in a function g of a random variable X, for example, antenna gain as a function of the E–W component of antenna pointing error. If the function g satisfies certain benign criteria (see Papoulis (1984)) then $g(X)$ is also a random variable. Its "expected value" $E(g(X))$ is defined by

$$E(g(X)) \triangleq \int_{\mathcal{S}} g(x) p_X(x) dx$$

The mean and variance of X can be expressed in this formulation, where $g(x) = x$ for the mean and $g(x) = (x - m_X)^2$ for the variance:

$$\begin{aligned} m_X &= E(X) \\ \sigma_X &= E\left[(X - m_X)^2\right] \end{aligned}$$

Suppose now we have two random variables, X and Y, each with any pdf. The "joint pdf" of X and Y is written $p_{X,Y}$ and is defined on all pairs of values of X, Y, respectively. X and Y are said to be **independent** if

$$p_{X,Y}(x, y) = p_X(x) p_Y(y) \text{ for all values } x \text{ of } X \text{ and } y \text{ of } Y$$

that is, if the probability that X has any particular value is not influenced by the value that Y has and vice versa. The concept of independence extends to more than two random variables, in which case the term "mutual independence" is more often used for clarity.

A concept related to independence or lack thereof is **correlation**. The amount of correlation between any two random variables X and Y is given by their correlation coefficient $\rho_{X,Y}$:

$$\rho_{X,Y} \triangleq \frac{E[(X - m_X)(Y - m_Y)]}{\sigma_X \sigma_Y}$$

Note that the correlation coefficient can be negative. The special case where $Y - m_Y = k(X - m_X)$ for some $k \neq 0$ provides us with the bounding values of the correlation coefficient, namely $+1$ and -1:

$$\rho_{X,Y} = \text{sgn}(k)$$

If X and Y are independent, they are uncorrelated. The inverse does not necessarily hold; however, if X and Y are Gaussian then it does hold.

8.A.4 Sum of Random Variables

We often have to deal with sums of random variables, for example, line items in performance budgets. Suppose X and Y are random variables. Then their sum $X + Y$

is also a random variable. Obviously, the sum of more than two random variables is also a random variable. The mean of the sum is just the sum of the individual means:

$$m_{X+Y} = m_X + m_Y$$

The standard deviation of the sum is more complicated. It depends on how correlated X and Y are:

$$\sigma_{X+Y} = \sqrt{\sigma_{X+Y}^2}$$
$$\sigma_{X+Y}^2 = \sigma_X^2 + \sigma_Y^2 + 2\rho_{X,Y}\sigma_X\sigma_Y$$

Examples for when $\sigma_X = \sigma_Y$ are as follows:

$$\sigma_{X+Y}^2 = \begin{cases} 4\sigma_X^2 & \text{when } \rho_{X,Y} = 1 \\ 2\sigma_X^2 & \text{when } \rho_{X,Y} = 0 \\ 0 & \text{when } \rho_{X,Y} = -1 \end{cases}$$

The sum of Gaussian random variables is Gaussian, but no other pdf family has this property (Papoulis, 1984).

8.A.5 Gaussian Probability Density Function

The Gaussian pdf is the most frequently used pdf in payload and communications system analysis. There are several reasons for this. One is that some variables truly do have a Gaussian pdf or very close to it. A second reason is that some variables have a distribution only partially known but known to be Gaussian-like in some respects. The best reason is that usually analysis is based on the sum of several random variables, and sums are usually Gaussian-like (see Section 13.2.5.1 on Central Limit Theorem). Sometimes sums are of many small terms of roughly equal size; they are of roughly equal size because if there were a term much larger than the others, it would be reduced by way of some design improvement. The last reason for the frequency of the Gaussian assumption is its ease of use. For example, the Fourier transform of a Gaussian pdf is a Gaussian curve, within a scaling factor. Computations with Gaussian pdfs can often be done analytically (i.e., by hand) or are at least simpler than with other pdfs, and estimates can sometimes be made quickly. Simulations are almost always faster because Monte Carlo can usually be avoided (Section 8.5). We have to caution here that a Gaussian pdf can only be assumed after careful assessment.

The Gaussian pdf corresponding to a mean m and a standard deviation σ is given by

$$p(x) = \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi}\sigma}$$

and is plotted in Figure 8.A.1 for mean of zero.

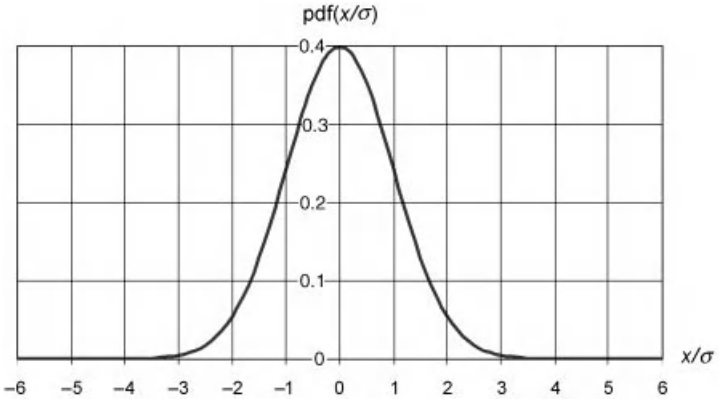


FIGURE 8.A.1 Gaussian pdf with zero mean.

The probabilities associated with 1σ and 2σ values are illustrated in Figure 8.A.2. Similar values for 3σ are the following: 99.68% for the probability that a Gaussian random variable is within $\pm 3\sigma$ of its mean and 99.83% for the probability that it is less than its mean plus 3σ, respectively.

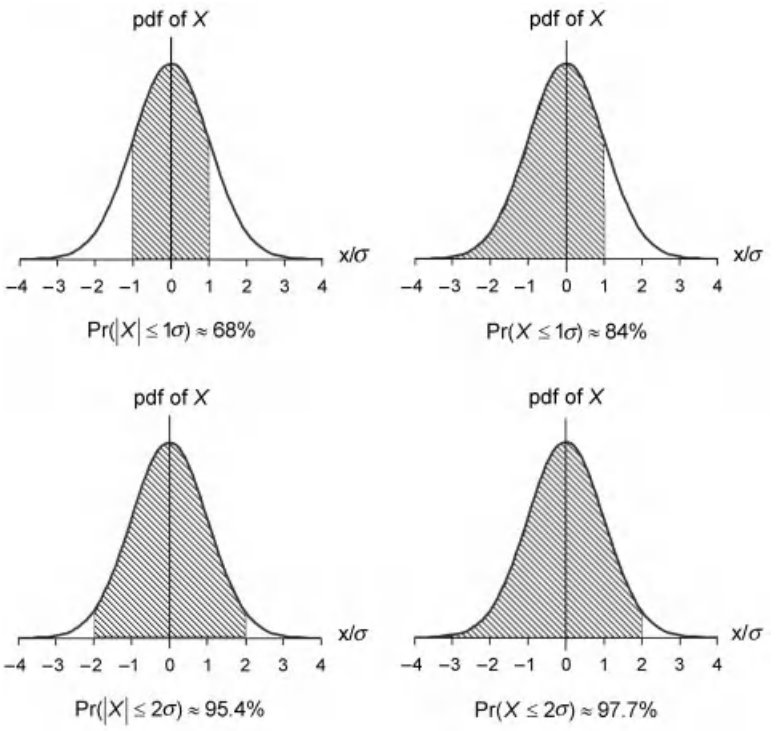


FIGURE 8.A.2 Some properties of Gaussian pdf with zero mean.

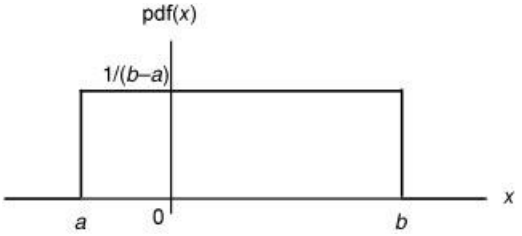


FIGURE 8.A.3 Uniform probability density function.

8.A.6 Uniform and Panel-Illumination Probability Density Function

Besides the Gaussian pdf, the most common pdf we meet is the **uniform pdf**, which has a constant value on some interval and is 0 outside. A less common pdf is for the diurnal variation in panel illumination (Section 8.3.2.7).

The uniform pdf usually comes in one of two forms: nonzero on a ±interval or nonzero from 0 to a positive number. The standard deviation for two such cases is given below, and the general pdf is plotted in Figure 8.A.3:

$$\text{Standard deviation for uniform density on} \begin{cases} -1 \text{ to } +1 \text{ is } 1/\sqrt{3} \approx 0.577 \\ 0 \text{ to } +1 \text{ is also } 1/\sqrt{3} \approx 0.577 \end{cases}$$

Another useful pdf is the **panel-illumination pdf**. Over the day, the earth deck is in the sun half the day (if it is not in the shadow of an antenna or support structure, anyway) and in the shade the other half. The amount of illumination starts at zero, acts like a sine function until it hits zero again, and then stays at zero for the second half of the day. The same is true for the east and west panels (refer to Figure 8.7). The diurnal illumination is plotted in Figure 8.A.4a, where the illumination phase is arbitrary. The diurnal illumination can be considered a random variable if the time at which we want to know what the illumination is, is random and uniformly

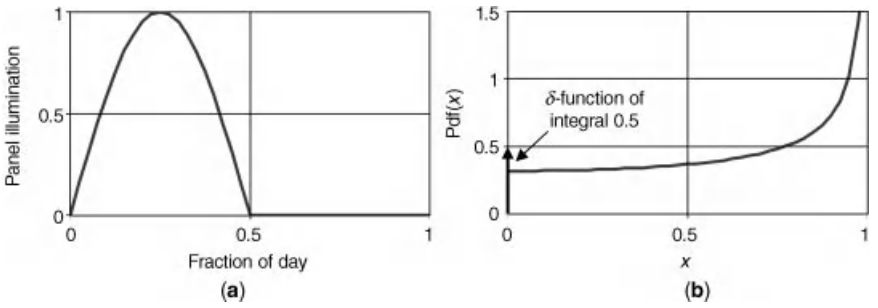


FIGURE 8.A.4 Diurnal variation in panel illumination: (a) illumination over the day and (b) pdf.

distributed over the day. The pdf for the diurnal illumination variation is drawn in Figure 8.A.4b and stated as follows:

Pdf for diurnal variation in panel illumination at x from 0 to 1 is $\frac{1}{2}\delta(x) + \left(\pi\sqrt{1-x^2}\right)^{-1}$ which has mean $\pi^{-1} \approx 0.318$ and std. dev. $\sigma \approx 0.386$

The corresponding random variable is a mixed-type (Appendix 8.A.2) since the distribution function is discontinuous at $x = 0$, jumping in value from 0 to 0.5. The use of this pdf is to approximate the diurnal temperature distribution of units on the inside of these panels and resultant variation of insertion loss. The pdf's independent axis values would range from diurnal minimum to maximum.

8.A.7 Standard Deviation of Drift of Unknown Magnitude and Direction

Suppose we have a random variable X which has a pdf that depends on the time z in the spacecraft life, where z is a fraction of the life. We wish to find the standard deviation of X averaged over life. First we have to find X's pdf. Let Z be the random variable for the time in the life. The joint pdf of X and Z is as follows (Papoulis, 1984):

$$p_{X,Z}(x,z) = p_{X|Z}(x|z)p_Z(z) = p_{X|Z}(x|z)$$

where the vertical bar means "conditioned on." The pdf of X is the "marginal pdf" $p_X(x)$:

$$p_X(x) = \int_0^1 p_{X,Z}(x,z)dz = \int_0^1 p_{X|Z}(x|z)dz$$

We further assume that X is a linear drift over life that at any given time z is uniformly distributed on the interval $[-z\alpha, +z\alpha]$ for $\alpha > 0$ (this is what variation from aging and radiation looks like). Then the pdf of X is as follows:

$$p_X(x) = -\frac{1}{2}\ln\left(\frac{|x|}{\alpha}\right) \text{ for } x \in [-\alpha, \alpha]$$

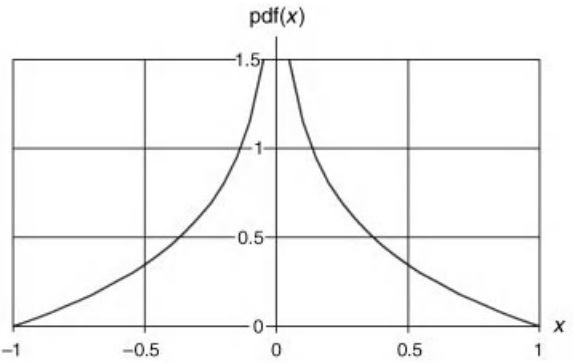


FIGURE 8.A.5 Pdf, averaged over life, of drift of unknown magnitude and direction.

which has mean 0 and standard deviation $\alpha/3$. The pdf is drawn in Figure 8.A.5 for $\alpha = 1$.

Now suppose we have another such drift Y, which at time z is uniformly distributed on the interval $[-z\beta, +z\beta]$ for $\beta > 0$. At any given z, X and Y are independent, but looked at over life they are not. However, they are uncorrelated. Therefore, from Appendix 8.A.4 we know that the standard deviation of the sum is the rss of the individual standard deviations.

If we want to look at the pdf at EOL and not averaged over life then the divisor in the standard deviation is not 3 but the square root of 3.

REFERENCES

Abramowitz M and Stegun IA, editors, (1965). *Handbook of Mathematical Functions, with Formulas, Graphs, and Mathematical Tables*. New York: Dover Publications.

Agilent Technologies, Inc (2009). Fundamentals of RF and microwave power measurements, part 3. Application note 1449-3. June 5.

Jones JJ (1963). Hard-limiting of two signals in random noise. *IEEE Transactions on Information Theory*; 9; 34–42.

Papoulis A (1984). *Probability, Random Variables, and Stochastic Processes*, International Student ed., 2nd ed. Singapore: McGraw-Hill, Inc.

Ramo S, Whinnery JR, and Van Duzer T (1984). *Fields and Waves in Communication Electronics*, 2nd ed. New York: John Wiley & Sons, Inc.

CHAPTER 9

PROCESSING PAYLOAD

9.1 INTRODUCTION

There are fewer processing payloads than nonprocessing, but the number is increasing as advancing technology allows the payload's mission to become more complex. A processing payload is a payload that has flexibility in most or all of the following: bandwidth allocation, gain, uplink-to-downlink channel mapping, channel-to-beam assignment, and power allocation. Some processing payloads do this with all-analog hardware but most have a digital processor. In terms of how deeply the payload processes the signal, there are two kinds of processing payloads, regenerative and nonregenerative, where the former performs demodulation and remodulation.

This chapter describes the processing payloads currently (2010) on orbit or about to be launched for which the author could find sufficient information. The payloads provide a service and do not simply demonstrate a possible future service. The list of such satellites changes about once a year as companies start up, stop a service, or go out of business.

There are a few reasons why a communications satellite may have a processing payload, among which are the following:

- Infeasibility of performing the quantity of desired processing with traditional hardware
- Accommodation of daily variation in traffic patterns
- Accommodation of long-term traffic changes
- Accommodation of satellite mission change, for example, orbital slot change
- Provision of direct user-to-user connectivity.

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

The rest of this chapter has the following sections:

- *Section 9.2*: Capabilities of current processing payloads
- *Section 9.3*: Digital-processing elements common to both nonregenerative and regenerative payloads
- *Section 9.4*: Nonregenerative processing payload
- *Section 9.5*: Regenerative payload.

9.2 CAPABILITIES OF CURRENT PROCESSING PAYLOADS

Let us see what the capabilities are of the current processing payloads. There is quite a variety of such payloads. The ones that the author could obtain sufficient information on are represented in Table 9.1 for nonregenerative ones and in Table 9.2 for regenerative ones. In each table, the payloads are ordered approximately by year of first launch. WGS is the US military's Wideband Global Satcom, formerly called the "Wideband Gapfiller Satellite." Some of the unrepresented ones are Iridium, Nilesat, and the US military's Milstar and Advanced EHF. All the satellites are GEO (except for Iridium). The tables show the capabilities of the part of each payload that has processing; this part includes any channel or beam switching that is outside of the processor but used with the processor. The quoted downlink bandwidth or bit rate includes the multiplying effect of any frequency reuse. "Channel" is in the frequency domain. "G/W" stands for **gateway**, which is a ground station that connects to terrestrial networks that serve as a resource for the users. Recall that in this book frequencies around 20 GHz are K-band, not Ka-band, so the uplink/downlink frequency combination of 30/20 GHz is Ka/K band.

Flexible bandwidth allocation is sometimes called "bandwidth on demand (BOD)," while flexible power allocation is sometimes called "power on demand (POD)." These capabilities are currently all managed on the ground, and the payload is commanded.

Five out of the six nonregenerative processing payloads connect users to gateways. Communication links to the user are said to be on the **forward link** and those from the user are on the **return link**, as shown in Figure 9.1. Communications via gateways is almost a defining characteristic of the nonregenerative processing payloads. As a consequence, only about half of the downlink traffic goes directly to users. The exception to the use of gateways is WGS, for which all communications are user-to-user, by request and scheduling. Common characteristics of the payloads are processing at IF and flexible bandwidth allocation, gain, uplink-to-downlink channel mapping, and power. Almost all have flexible channel-to-beam assignment. The only two categories in which the choice is really open are the following: (1) channel filtering and routing technology, in which half of the payloads use analog and half use digital; and (2) beam-forming method, in which only two use digital beam-forming with a radiating array. Analog technology is used on both older and newer payloads. Digital beam-forming is only used at the relatively low

TABLE 9.1 Capabilities of Some Nonregenerative Processing Payloads

Property/Satellite(s)	Inmarsat-3	Inmarsat-4	Thuraya	Anik F2	WGS	Hylas
Processor or payload name	–	–	–	Beam link	–	Generic flexible payload
Successful launches	1996–98	2005–08	2003, 2008	2004	2007–12 (planned)	2010
Frequency band(s)	L- and C-bands	L- and C-bands	L- and C-bands	Ka/K-band	X- and Ka/K-bands	Ka/K-band
Downlink BW or bit rate per satellite, processed	63.5 MHz	272 MHz	1 GHz	3 GHz	3.6 Gbps	1.3 GHz
User connectivity	User–G/W–user	User–G/W–user	User–G/W–user	User–G/W–user	User–user, user–G/W, G/W–user	G/W–user, user–G/W
Frequency at which processing occurs	160 MHz	IF	IF	L-band	IF	C-band
Flexible bandwidth allocation	Yes	Yes	Yes	Yes	Yes	Yes
Flexible gain	Yes	Yes	Yes	Yes	Yes	Yes
Flexible uplink-to-downlink channel mapping	Yes	Yes	Yes	Yes	Yes	Yes
Flexible channel-to-beam assignment	Yes	Yes	Yes	Yes	Yes	No
Channel filters and routing technology	Analog	Digital	Digital	Analog	Digital	Analog
Flexible power allocation	Yes	Yes	Yes	Yes	Yes	Yes
Digital beam-forming	No	Yes	Yes	No	No	No

Sources: Peach et al. (1994), International Technology (1998), Martin et al. (2007), Farrugia (2006), EADS (2009), Thuraya (2009a, 2009b), Sunderland et al. (2002), Lee et al. (2002), Boeing (2004), Hong et al. (2008), Kumar et al. (2005), Bennett et al. (2006), Nguyen et al. (2003), Anzalchi and Harverson (2007), Avanti (2009), ESA (2006).

TABLE 9.2 Capabilities of Some Regenerative Payloads

Property/Satellite(s)	Hot Bird 6	W3A	Amazonas	Spaceway-3
Processor(s) or payload name	SkyPlex	SkyPlex and Enhanced Skyplex	AmerHis	–
Successful launches	2002	2004	2004, 2009	2007
Frequency band(s)	Ku- and K-bands	Ku- and K-bands	Ku-band	Ka/K-band
Downlink BW or bit rate per satellite, processed	330 Mbps	275 Mbps	216 Mbps	10 Gbps
User connectivity	G/W–user	G/W–user	User–user, G/W–user, user–G/W	User–user, G/W–user, user–G/W
Flexible bandwidth allocation	On uplink	On uplink	Yes	Yes
Flexible gain	Yes	Yes	No	No
Flexible uplink-to-downlink channel mapping	Yes	Yes	Yes	Yes
Flexible channel-to-beam assignment	No	No	Yes	Yes
Channel filters and routing technology	Analog	Digital filters on Enhanced Skyplex	Digital	Digital
Flexible power allocation	No	No	No	Yes
Digital beam-forming	No	No	No	No
Decoding	Some proc units, yes, turbo	Yes, turbo	Yes, turbo	Yes
Coding	Yes	Yes	Yes	Yes
Communications format	DVB-S	DVB-S	RSM-B	RSM-A

Sources: Eutelsat (2010a), Wittig (2003), Novello and Piloni (2000), ESA (2004a, 2004b), Eutelsat (2010b), Novello and Piloni (2000), Wittig (2003), ETSI (2004, 2006, 2009), Thales (2009), Whitefield et al. (2006), Hughes (2008), Wu et al. (2003).

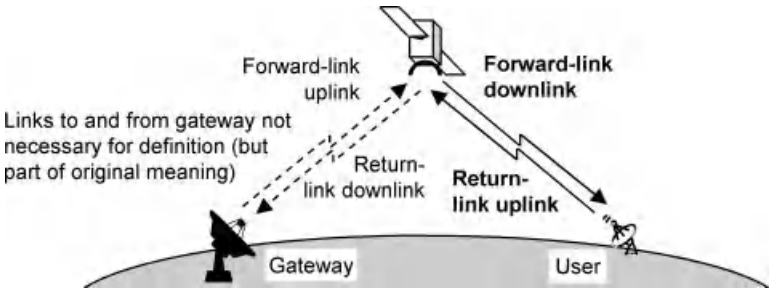


FIGURE 9.1 Definition of forward and return links.

frequencies of L- and C-bands. The processing hardware in these payloads will be discussed in the following two sections.

A defining characteristic of the regenerative payloads is not, as you might expect, that they connect users directly to each other without gateways. Only two of the four regenerative payloads provide that; the other two, the Eutelsat satellites, combine TV signals from multiple TV broadcasting centers and broadcast them to consumers. Characteristics of the regenerative payloads are that all have flexible uplink-to-downlink channel mapping, none have digital beam-forming, and almost all provide error-correction decoding of the uplink. The Amazonas and Spaceway-3 payloads do not provide flexible gain control, but their systems instead provide the means for the uplinking terminals to adjust their output power. Some of the hardware used in regenerative payloads is discussed in the following section, and these payloads themselves are discussed in Section 9.5.

9.3 DIGITAL-PROCESSING ELEMENTS COMMON TO BOTH NONREGENERATIVE AND REGENERATIVE PAYLOADS

The hardware elements discussed in this section may be found in any payload that performs digital processing, namely some nonregenerative payloads and all regenerative payloads. Units unique to the two kinds of processing payload are described in the two sections that follow this one.

Digital processing has the advantage of increasing the payload capability, to a scale not possible with analog technology. For example, the payload could handle more spot beams, hundreds of them as on Spaceway-3, allowing smaller beams and thus higher frequency reuse, which means increased capacity. Digital processing currently has the disadvantages of high size, mass, and power requirements and of the risk associated with technology adolescence (Chie, 2010).

9.3.1 A-to-D Converter

Digital processing elements in a payload must be preceded by an **analog-to-digital converter** (**ADC**, **A-to-D**, or **A/D**). A block diagram of an ADC and associated elements is shown in Figure 9.2, along with an example of quantizer levels. The number of levels in an actual ADC would be greater.

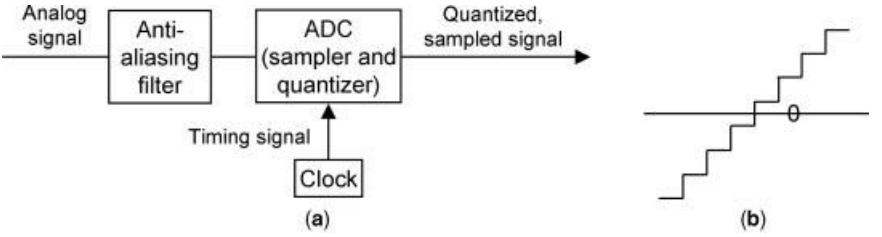


FIGURE 9.2 A-to-D converter: (a) ADC with associated elements and (b) levels of 3-bit quantizer.

The ADC must sample at least as fast as the **Nyquist rate** of the signal, that is, twice the highest frequency of the signal. This is because the sampling will cause any frequency components beyond $\pm 1/2$ sampling rate to overlay the components within $\pm 1/2$ the sampling rate (Section 10.5.3.2). So any unwanted high-frequency components must be filtered out before the sampling or they will distort the sampler output. The filter for this is the **anti-aliasing** filter. This filter and the clock are often not shown in diagrams but they are understood to be present.

The next element in the ADC is the sampler-and-quantizer. A clock provides the timing signal to this. The sampler outputs a weighted average of its input over a period of time. The quantizer has a stated number of levels, which is a power of two, where the power is the number of bits. The quantizer will clip any value that lies outside its min/max.

A SNR can be defined for the ADC, the ratio of signal power to power of the added noise within the signal bandwidth. The ADC introduces noise to the input signal, like analog components do, but the main noise sources are different. One type of noise is quantization noise, which is uniformly distributed over $\pm 1/2$ quantization step size. The other main source of noise, for sampling rates from about 2 Ms/s to about 4 Gs/s, is aperture jitter. This is a timing jitter: the sampling epoch is characterized by a mean and a standard deviation. For even higher sampling rates, the main noise source besides quantization noise is comparator ambiguity, which comes from the comparator responding too slowly to a small voltage difference from last sample to this. The occurrence is related to the comparator's regeneration time (Walden, 1999a, 1999b).

An effective number of bits can be defined, which is less than the stated number of bits by a reduction that accounts for ADC noise other than quantization error. The reduction is about 1.5 bits for all sampling rates. A universal measure of ADC performance is the product of its effective number of quantization levels and its sampling rate (Walden, 1999b).

In regard to the signal power, if clipping rarely comes into play then the input and output signal power are the same. However, if the input signal either has a lot of noise on it already or is noise-like itself (e.g., consists of many equal-magnitude carriers), the ADC will clip and thus create intermodulation products (IMPs) (Section 5.3). Some signal power is lost to the IMPs, which are treated as an additional output noise term (Section 13.2.4). The IMPs are minimized if the quantizer levels are centered about zero as shown in Figure 9.2b (Taggart et al., 2007).

9.3.2 D-to-A Converter

Digital processing elements in the payload must be followed by a **digital-to-analog converter** (**DAC**, **D-to-A**, or **D/A**). The DAC turns the digital signal into an analog signal while using the timing clock to do so. This analog signal has unwanted high-frequency components that are removed by the following **reconstruction filter** (Salim et al., 2004). This process is illustrated in Figure 9.3. The reconstruction filter adds noise to the signal like all analog elements do. This filter and the clock are often not shown in diagrams but they are understood to be present.

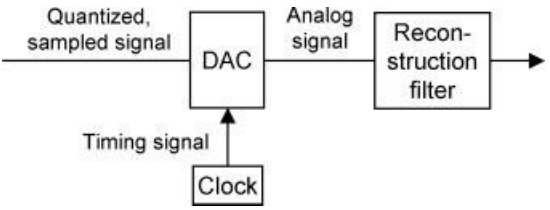


FIGURE 9.3 D-to-A converter with associated elements.

9.3.3 Digital Filtering

A **polyphase filter** is a parallel-processing implementation of a **finite-impulse-response (FIR)** filter. Computations with a FIR filter are done in the time domain, since the impulse response has only a finite number of samples. Interpolators and decimators are examples of polyphase filters (Purcell, 2010). Interpolators increase the sampling rate of a signal and decimators decrease it, as we will see shortly. Interpolators are used in digital demodulation and digital beam-forming (Pellon, 1992; Godara, 1997).

A general technique of signal processing is **multirate filtering**. It is a computationally efficient way to implement a FIR filter that has an extremely long impulse response. Different sampling rates are used at different steps along the way: decimation at the beginning and interpolation at the end (Purcell, 2010). Multirate filtering is used in payload processors.

Interpolation multiplies the sampling rate of a signal. It must work in a way so as not to add frequency components to the signal that were not present to begin with. Interpolation has two steps, as illustrated in Figure 9.4a. Suppose that the sampling rate is to be multiplied by an integer L. The first step is to pad the original samples with L-1 zeros after each original sample. This creates a signal with very sharp transitions, namely very high frequencies. The second step is to low-pass filter the signal. This **smoothing filter** performs smoothing that is consistent with the frequency components present in the original signal (Purcell, 2010).

Decimation divides the sampling rate of a signal. The usual assumption is that the sampling rate is higher than it needs to be to capture the frequency components present in the signal. Decimation has two steps, as illustrated in Figure 9.4b. Suppose that the sampling rate is to be divided by an integer M. The first step is to low-pass filter the signal, with an anti-aliasing filter. The second step is to delete M-1 samples out of every M (Purcell, 2010).

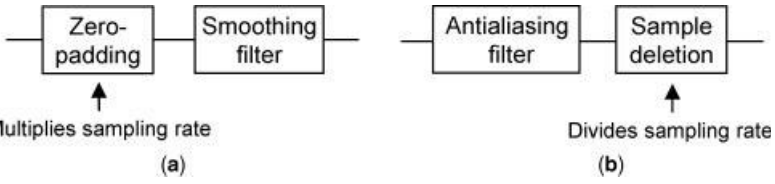


FIGURE 9.4 Two digital processing operations: (a) interpolation and (b) decimation.

9.4 NONREGENERATIVE PROCESSING-PAYLOAD

9.4.1 Payload Architecture

Nonregenerative processing payloads are sometimes called "transparent" or "bent-pipe" processing payloads in the literature. At the highest level the architecture of a nonregenerative processing payload is the same as that of a nonprocessing payload, but the names of some unit banks are different: the IMUX and/or OMUX banks are called the **channelizer**, the channel and beam-routing switch banks are called the **router**, and if there is an antenna array with onboard beam-former the latter may be a digital beam-former. There may be hundreds of channel and beam combinations. The router provides the connections of input beams and channels to output channels and beams (Cherkaoui and Glavac, 2008). In some cases the functions of channelizer and router are intertwined so the two units are combined. The three new kinds of units perform the same functions as in nonprocessing payloads but provide so much more flexibility that they require new unit architectures and technologies.

9.4.2 Analog Channelizer and Router

Hylas, Anik F2, and Inmarsat-3 have analog channelizers and routers.

The processing payload on Hylas has the **select-then-filter** architecture, which minimizes the number of channel filters to one per channel. It is often a good architecture when the overall channel frequency plan is regularly spaced and the number of channels used at any one time is relatively small, for example, half, compared to the number of possible channels that can be used. An example of this architecture is shown in Figure 9.5. The full frequency band is divided into subbands, and within every subband the channel spacings are the same (Anzalchi and Harverson, 2007).

The processing payload on Anik F2 employs the **switch-then-filter** architecture (Lee et al., 2002). (By the way, the "F2" in Anik F2 does not mean "flight 2" as is usual but means the second flight of the F series.) This architecture appears to be similar to the select-then-filter architecture (Lee et al., 2002).

The analog channel filters on Anik F2 and Inmarsat-3 are implemented in **surface acoustic-wave (SAW)** technology (Lee et al., 2002; Peach et al., 1994). Inmarsat-3 uses the sharp-roll-off property of these filters in a technique called **bandwidth-switchable SAW filtering**, to implement flexible bandwidth and guard-band recovery. Normally, there is a guard band between adjacent channel

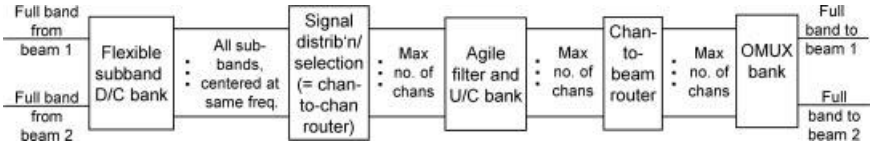


FIGURE 9.5 Example of select-then-filter analog processing-repeater architecture (after Anzalchi and Harverson (2007)).

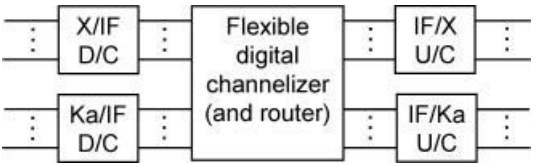


FIGURE 9.6 Partial architecture of WGS processing repeater (after Boeing (2009)).

bands, with bandwidth about 10% of a channel band, unused except to allow for a small overlap of the signal spectra and the detection filter responses. The SAW filters in a bank can be used singly or in combinations, where a combination passes the combined passbands including the guard bands (Peach et al., 1994).

A bandpass SAW filter consists of two transducers on a substrate, for input and output, which convert between voltage and mechanical stress. Between the transducers travels a surface acoustic wave. Because of the slow wave-velocity, long delays can be implemented in a small space (Com Dev, 2010). The kind of SAW filter used for channelizing is a "transversal filter," for its sharp roll-off (Norspace, 2011). A transversal filter's output is the sum of variously delayed and scaled versions of the input signal. The filter's inband ripple can be as low as 0.2 dB pk-pk, but the insertion loss is on the order of 20 dB (Com Dev, 2010).

9.4.3 Digital Channelizer and Router

Thuraya and WGS have digital channelizers and routers. Digital channelization can be performed in either the time or frequency domain. In the time domain the polyphase filter approach (Section 9.3.3) is used (Kumar et al., 2005). The reason is that the FIR filters are very long due to the corresponding frequency-domain filters being very sharp. If the channelization is performed in the frequency domain, the operation is preceded by a discrete Fourier transform (DFT) and followed by an inverse DFT.

Thuraya, which has 640 gateway channels in either direction, performs channelization in the time domain. Then it digitally routes any of the gateway channels to any of 219 mobile subbands and vice versa (Sunderland et al., 2002). It is impossible to imagine doing this with analog hardware.

The WGS satellites provide connectivity among subchannels and beams and even frequency bands (Kumar et al., 2005). Figure 9.6 illustrates this. There is considerable flexibility in the routing of the 1872 subchannels (Kumar et al., 2005).

9.4.4 Digital Beam-Former

Digital beam-forming is currently done only by nonregenerative payloads. A digital beam-former is a digital beam-forming network for a radiating array that performs all its computations digitally. On receive it samples, digitizes, and stores the signals from all the array elements. To form each beam, it delays the signals appropriately (instead of phasing), weights them, and combines them. On transmit for

each beam it samples and digitizes the corresponding signal from the repeater. For each array element, it delays and weights the signal properly, and it accumulates these numbers over all the beams (Godara, 1997). The use of delay instead of phase shift makes the antenna more wideband, as scanning causes no squint (Section 3.4). Beam-forming in general includes the function of beam-to-channel assignment on receive and channel-to-beam assignment on transmit, for antenna arrays (Craig et al., 1992).

All radiating arrays known to the author are flat and have the radiating elements in a regular grid pattern. Uniform spacing allows one to take full advantage of signal processing techniques, which are based on regularly spaced sampling (Dudgeon, 1977). Parallel processing is facilitated.

Digital beam-forming does not have to be executed directly at RF. The RF signals could be downconverted to an IF, decimated, processed at IF, interpolated, and upconverted back to RF (Section 9.3.3) (Purcell, 2010). Digital beam-forming is currently done only at L- and C-bands (see Tables 9.1 and 9.2), perhaps because the extra steps would be necessary at higher frequencies.

9.5 REGENERATIVE PAYLOAD

9.5.1 Introduction

A regenerative payload at the very least demodulates and remodulates the signal (for what these actions are, see Chapter 10). This has the possible disadvantage of freezing the uplink and downlink modulation schemes, that is, how the frequency band is used (Chie, 2010). However, it creates some advantages over a nonregenerative payload:

- Possibility of onboard combining of signals, thereby reducing or removing the need for terrestrial gateways
- Possibility of employing time-domain multiplexing (TDM), in which signals are assigned time slots in one modulated signal, so if TWTAs are used (and not SSPAs) they can be operated at their most power-efficient, maximizing use of satellite resources
- Possibility of onboard routing based on message headers in signals
- Improved end-to-end (ground-to-satellite-to-ground) signal-to-noise ratio (SNR), up to 3 dB but usually less.

If the payload does further processing, it can provide more advantages:

- User-to-user connectivity, which means about half the latency and about twice the use of uplink-plus-downlink data rate
- Error-correction decoding of uplink and recoding of downlink, giving much improved uplink SNR, which much reduces the capabilities required of the user transmitter and payload receiver.

The role that regenerative payloads play in their end-to-end communications systems is essentially different from the role of nonprocessing payloads and their conceptual equivalents, the nonregenerative payload. The regenerative payloads can only be understood in the context of their whole systems. The regenerative payloads summarized in Table 9.2 are of two kinds. The simpler kind, represented by Eutelsat's Hot Bird 6 and W3A, combine TV programs from broadcasting centers and broadcast them. The much more complex kind, represented by Amazonas and Spaceway-3, provide Ethernet-like user-to-user communications.

9.5.2 Current Regenerative Payloads in TV Broadcast Network

Eutelsat's Hot Bird 6 and W3A satellites broadcast TV in regional beams. In a system with a traditional payload, a hub or gateway would combine the TV programs from various broadcast centers for the payload. But these satellites are the hub, that is, they do the combining, as illustrated in Figure 9.7.

The satellites use one standard in the **Digital Video Broadcasting (DVB)** family of communications standards established by the European Telecommunications Standards Institute (ETSI). The DVB standard that the satellites use is **DVB-Satellite (DVB-S)** (Novello and Piloni, 2000), which specifies what kind of signal must be transmitted on the forward link but does not specify what the payload must do (EN, 1997). Each TV broadcast center combines various TV programs into a stream of **MPEG-2** packets, each of which consists of a header and a "payload" of useful data. MPEG denotes the Motion Picture Experts Group. The stream of MPEG-2 packets is the MPEG-2 **transport stream (TS)**, as shown in Figure 9.8 (EN, 1997). Each broadcast center digitally processes, modulates, and transmits its TS to the satellite (Wittig, 2003). The satellite payload combines the TSs from the various broadcast centers into one TS. It has to do this at baseband so it can recognize the structure of the MPEG-2 packets.

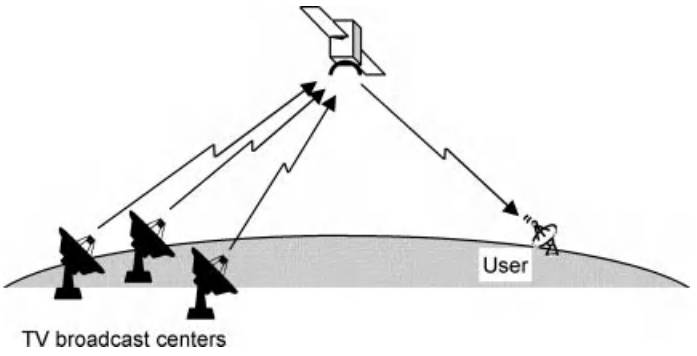


FIGURE 9.7 Communications connectivity provided by Eutelsat's Hot Bird 6 and W3A payloads (after Wittig (2003)).

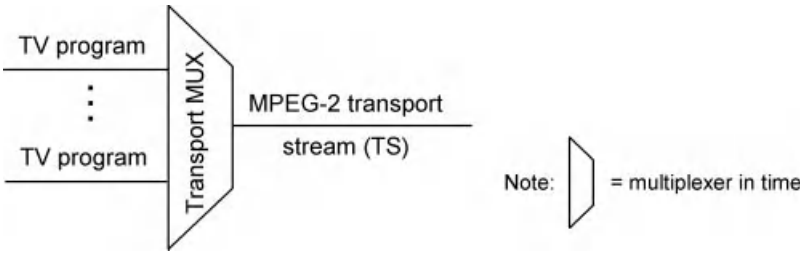


FIGURE 9.8 TV programs multiplexed into a MPEG-2 transport stream, in a TV broadcast center (after EN (1997)).

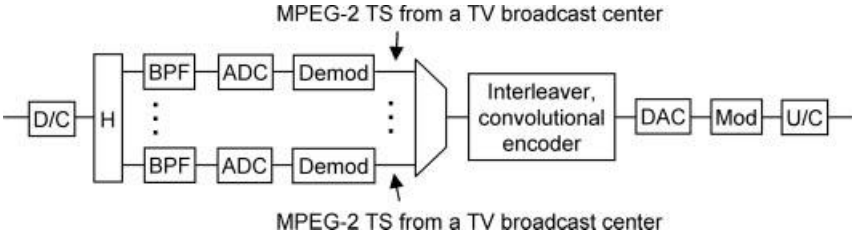


FIGURE 9.9 Eutelsat's second-generation Skyplex regenerative processor (after Wittig (2003)).

Hot Bird 6 and W3A each carry a few Skyplex processors. Hot Bird 6 has three of the second-generation processors and five of the updated second-generation processors (ESA, 2004b). Figure 9.9 shows a block diagram of the nonupdated second-generation processor. It performs no decoding. W3A has five of the updated second-generation and one of the enhanced Skyplex (ESA, 2004c); these two versions have the same block diagram (ESA, 2004b), shown in Figure 9.10. They perform turbo decoding, scrambling, interleaving, and convolutional encoding. Both versions of the second-generation processor employ SAW technology for the BPFs, but the enhanced processor employs digital technology instead (ESA, 2004b).

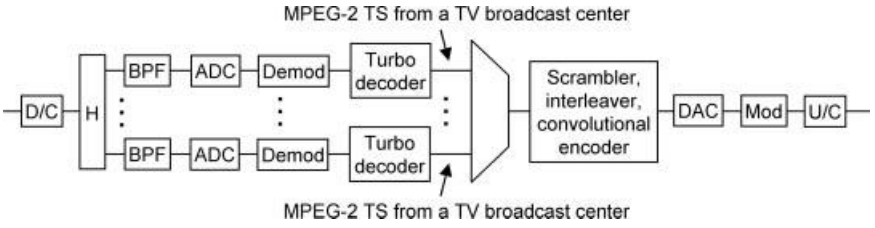


FIGURE 9.10 Skyplex processor, both updated second-generation and enhanced versions (after Wittig (2003)).

9.5.3 Current Regenerative Payloads in Mesh Network

Amazonas and Spaceway-3 both provide a **mesh network** with direct user-to-user connectivity, as well as a **star network** where users connect with gateways. The Amazonas and Spaceway-3 systems are similar in concept but different in scale. Their gateways provide connectivity to the Internet and other terrestrial networks, but the gateways have the same kind of terminal as the users do to communicate with the payload.

Amazonas and Spaceway-3 provide return links in addition to forward links. Both Amazonas and Spaceway-3 follow a **Regenerative Satellite Mesh (RSM)** standard. Amazonas uses the newer version, RSM-B (ETSI, 2009), which is a specialization of the **DVB–Return Channel Satellite (DVB–RCS)** standard (ETSI, 2006). This standard integrates DVB-S for the forward link and specifies the return link. Amazonas converts return-link signals in the specified return-link format into forward-link signals in DVB-S. Spaceway-3 uses an earlier version of RSM, RSM-A (Whitefield et al., 2006), which is not compliant with DVB (ETSI, 2004).

A user terminal in Amazonas has an **Internet Protocol (IP)** entity in it, which can receive and transmit **IP datagrams** (IP traffic consists of IP datagrams (Tanenbaum, 1996)). Suppose that the user terminal has some IP datagrams to transmit to another user terminal in the system. It performs **multiprotocol encapsulation (MPE)** on them, creating a transport stream (TS) of MPEG-2 transport packets (ETSI, 2006), as shown in Figure 9.11. The encapsulation segments the IP datagrams and adds a header to each segment. The MPEG-2 packets exist in the **medium-access control (MAC)** sublayer of computer communications (Tanenbaum, 1996). The best-known MAC is the Ethernet. The terminals have MAC addresses, and the payload switches the packets according to their destination MAC address. The receiving terminal receives the MPEG-2 packets and undoes the encapsulation, returning them to IP datagrams, and passes them to its IP entity. Basically the payload and the user terminals form an Ethernet in the sky. Spaceway-3 is similar except that the encapsulation is a different kind from MPE.

The Amazonas system uses multiple frequencies for a connection in addition to TDM (ETSI, 2006), so the payload is complex. Figure 9.12 offers a diagram of the AmerHis regenerative repeater onboard Amazonas.

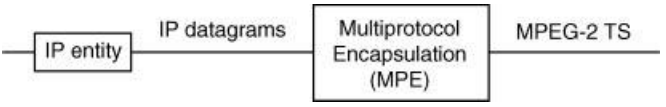


FIGURE 9.11 IP datagram flow encapsulated into a MPEG-2 transport stream, in Amazonas user terminal.

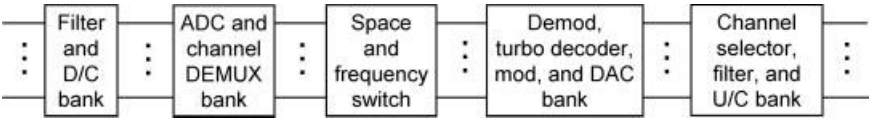


FIGURE 9.12 Amazonas' regenerative repeater AmerHis (after Winkler et al. (2009)).

Spaceway-3 has a huge number of downlink cells, 784, that it connects with 112 uplink cells (Hughes, 2009). The cells are not quite spot beams because they do not all exist simultaneously. The payload makes the connections in a real-time fashion by transmitting on so-called hopping spot beams (Boeing, 2007), discussed in Section 3.5.6. The 24 simultaneous downlink spot beams each hop consecutively and extremely rapidly among 32 cells (Hughes, 2009). The payload must use a digital processor to perform its electronic beam-steering.

REFERENCES

Anzalchi J and Harverson M (2007). Generic flexible payload technology for enhancing in-orbit satellite payload flexibility. *AIAA International Communications Satellite Systems Conference;* Apr. 10–13; 1–15.

Avanti Communications Group PLC. Hylas. On www.avantiplc.com. Accessed Oct. 2009.

Bennett B, Hannan D, Marshall J, and Gibbons R (2006). Link analysis of commercial and Wideband Gapfiller Satellite (WGS) satellites using DVB-S2 with variable coding and modulation (VCM). *IEEE Military Communications Conference;* Oct. 23–25; 1–7.

Boeing Defense, Space & Security (2004). Anik F2, Ka-band for North America. Program fact sheet. On www.boeing.com/defense-space/space/bss/programs.html. Accessed Jan. 2010.

Boeing Defense, Space & Security (2007). Spaceway 3 North America, bandwidth on demand. Program fact sheet. On www.boeing.com/defense-space/space/bss/programs.html. Accessed Jan. 2009.

Boeing Defense, Space & Security (2009). Transformational wideband communication capabilities for the warfighter. WGS program fact sheet. On www.boeing.com/defense-space/space/bss/programs_nscp.html. Accessed Jan. 2010.

Cherkaoui J and Glavac V (2008). Signal frequency channelizer/synthesizer. *International Workshop on Signal Processing for Space Applications;* Oct. 6–8; 1–4.

Chie CM, retired from Boeing Satellite Center (2010). Personal communication; Mar. 7.

Com Dev International Ltd (2010) Com Dev SAW filters. Application note 102. On www.comdev.ca/docs/com_dev_saw_filters.pdf. Accessed Jan. 27.

Craig AD, Leong CK, and Wishart A (1992). Digital signal processing in communications satellite payloads. *IEE Electronics & Communication Engineering Journal;* June; 107–114.

Dudgeon DE (1977). Fundamentals of digital array processing. *Proceedings of IEEE;* 65; 898–904.

EADS, Astrium (2009). Inmarsat-4, the very latest in communications technology. Program article. On www.astrium.eads.net. Accessed 2009 Mar.

EN 300 421, European Std. (Telecommunications series), v1.1.2 (1997). Digital video broadcasting; framing structure, channel coding and modulation for 11/12 GHz satellite services. Joint Technical Committee of the European Broadcasting Union, Comité Européen de Normalisation Electrotechnique, and European Telecommunications Standards Institute.

ESA (2004a). AmerHis, a new generation of satellite communications systems. Publication BR-226. The Netherlands: ESA Publications Division; June.

ESA (2004b). Enhanced Sykplex EQM passes testing. Article. On telecom.esa.int. First written July 28, 2003; updated Feb. 17, 2004.

ESA (2004c). SkyPlex: flexible digital satellite telecommunications. Article. On www.esa.int. Mar. 9.

ESA (2006). Contract for the innovative flexible broadband satellite. Hylas. Article. On www.esa.int. May 15.

ETSI, Technical Spec (TS) 102 188-1, v1.1.2 (2004). Satellite earth stations and systems (SES); regenerative satellite mesh—A (RSM-A) air interface; physical layer specification; part 1: general description. Sophia-Antipolis Cedex, France: European Telecommunications Standards Institute.

ETSI, TS 102 429-1, v1.1.1 (2006). Satellite earth stations and systems (SES); broadband satellite multimedia (BSM); regenerative satellite mesh—B (RSM-B); DVB-S/DVB-RCS family for regenerative satellites; part 1: system overview. Sophia-Antipolis Cedex, France: European Telecommunications Standards Institute.

ETSI, TS 102 602, v1.1.1 (2009). Satellite earth stations and systems (SES); broadband satellite multimedia; connection control protocol (C2P) for DVB-RCS; specifications. Sophia-Antipolis Cedex, France: European Telecommunications Standards Institute.

Eutelsat (2010a). Hot Bird 6. Brochure. On www.eutelsat.com. Accessed Feb. 2010.

Eutelsat (2010b). W3A. Brochure. On www.eutelsat.com. Accessed Feb. 2010.

Farrugia L, EADS Astrium UK (2006). Astrium view of future needs for interconnect complexity of telecommunications satellite on-board digital signal processors. Presentation at meeting at European Space Research and Technology Centre; Feb. 9. On escies.org/GetFile?rsrcid=1643. Accessed Feb. 1, 2010.

Godara LC (1997). Application of antenna arrays to mobile communications, part II: beamforming and direction-of-arrival considerations. *Proceedings of the IEEE;* 85; 1195–1245.

Hong Y, Srinivasan A, Cheng B, Hartman L, and Andreadis P (2008). Optimal power allocation for multiple beam satellite systems. *IEEE Radio and Wireless Symposium;* Jan. 22–24; 823–826.

Hughes Network Systems, LLC (2008). Network connectivity on Spaceway. White paper. On www.hughes.com. October. Accessed Mar. 19, 2009.

Hughes Network Systems, LLC (2009). Spaceway 3 technical specifications. On www.hughes.com. October. Accessed Mar. 10, 2010.

International Technology Research Institute, World Technology Division (1998). Site report on Com Dev. Final report, World Technology (WTEC) panel on global satellite communications technology and systems, Appendix E. Dec. On www.wtec.org/loyola/pdf/satcom2.pdf. Accessed Apr. 14, 2010.

Kumar R, Taggart D, Monzingo R, and Goo G (2005). Wideband Gapfiller Satellite (WGS) system. *IEEE Aerospace Conference;* Mar. 5–12; 1410–1417.

Lee M, Wright S, Dorey J, King J, and Miyakawa RH (2002). Advanced Beam Link processor for commercial communication satellite payload application. *AIAA International Communications Satellite Systems Conference;* May 12–15; 1–11.

Martin DH, Anderson PR, and Bartamian L (2007). *Communications Satellites*, 5th ed. El Segundo, CA; Reston, VA: The Aerospace Press; American Institute of Aeronautics and Astronautics, Inc.

Nguyen LN, Kinal GV, and Paul HI (2003). Performance of hybrid link power control using Wideband Gapfiller Satellite-based automatic level control (ALC) and ground-based link power control (LPC). *IEEE Military Communications Conference;* 1 (Oct. 13–16); 511–516.

Norspace (2011). SAW filters. Product information. On www.norspace.no/artikkel.html?sider.sid=6&kid=5&mp=2. Accessed Sept. 12.

Novello R and Piloni V (2000). Skyplex: its success and evolution towards broadband multi-spot OBP. *ESA Millennium Conference on Antennas and Propagation;* Apr. 9–14; 1–7.

Peach RC, Lee YM, Miller ND, van Osch B, Veenstra A, Kenyon P, and Swarup A (1994). The design and implementation of the Inmarsat 3 L-band processor. *AIAA International Communications Satellite Systems Conference;* Feb. 27–Mar. 3; 1–11.

Pellon LE (1992). A double Nyquist digital product detector for quadrature sampling. *IEEE Transactions on Signal Processing;* 40; 1670–1680.

Purcell JE, Momentum Data Systems Inc (2010). Multirate filter design—an introduction. Application note. On www.mds.com. Accessed 2010 Feb.

Salim T, Devlin J, and Whittington J (2004). Analog conversion for FPGA implementation of the TIGER transmitter using a 14 bit DAC. *Proceedings of IEEE International Workshop on Electronic Design, Test and Applications;* Jan. 28–30; 437–439.

Sunderland DA, Duncan GL, Rasmussen BJ, Nichols HE, Kain DT, Lee LC, Clebowicz BA, and Hollis IV RW (2002). Megagate ASICs for the Thuraya satellite digital signal processor. *Proceedings of IEEE International Symposium on Quality Electronic Design;* Mar. 18–21; 479–486.

Taggart D, Kumar R, Krikorian Y, Goo G, Chen J, Martinez R, Tam T, and Serhal E (2007). Analog-to-digital converter loading analysis considerations for satellite communications systems. *IEEE Aerospace Conference*, Mar. 3–10; 1–6.

Tanenbaum AS (1996). *Computer Networks*, 3rd ed. New Jersey: Prentice Hall PTR.

Thales Alenia Space (2009). Amazonas-2 satellite to embark Thales Alenia Space's Amerhis-2 system. Press release. Sept. 22. On www.thalesgroup.com/Pages/PressRelease.aspx?id=10303. Accessed Jan. 26, 2010.

Thuraya Telecommunications Co (2009a). Space segment. On www.thuraya.com. Accessed Jan. 27, 2010.

Thuraya Telecommunications Co (2009b). Thuraya satellite. On www.thuraya.com. Accessed Mar. 19, 2009.

Walden RH (1999a). Analog-to-digital converter survey and analysis. *IEEE Journal on Selected Areas in Communications;* 17; 539–550.

Walden RH (1999b). Performance trends for analog-to-digital converters. *IEEE Communications Magazine;* Feb.; 96–99.

Whitefield D, Gopal R, and Arnold S (2006). Spaceway now and in the future: On-board IP packet switching satellite communication network. *IEEE Military Communications Conference;* Oct. 23–25; 1–7.

Winkler R, Losquadro G, and Mura R, inventors; Finmeccanica SpA, assignee (2009). System for satellite digital communications with onboard processing adaptive to the attenuation traffic and radio interface conditions for point to point communication and meshed grids. U.S. patent 2009/0303916 A1. Dec. 10.

Wittig M, European Space Research and Technology Centre (2003). Telecommunikation satellites: the actual situation and potential future developments. Presentation; Mar. On www.dlr.de/rd/Portaldata/28/Resources/dokumente/RK/wittig-multimedia_systeme.pdf. Accessed Jan. 28, 2010.

Wu YA, Chang RY, and Li RK (2003). Precision beacon-assisted attitude control for Spaceway. *AIAA Guidance, Navigation, and Control Conference;* Aug. 11–14.

PART II

PAYLOAD IN END-TO-END COMMUNICATIONS SYSTEM

CHAPTER 10

PRINCIPLES OF DIGITAL COMMUNICATIONS THEORY

10.1 INTRODUCTION

This chapter explains the basic concepts of digital communications theory for the following purposes:

- To show why the payload communications parameters of Chapter 7 matter
- To serve as background for the rest of the book, which treats the payload set in its place in the end-to-end communications system.

The chapter contains more drawings than equations, so it can be as useful as possible to working engineers. For equations not all the validity prerequisites are stated, since in every normal case of satellite communications the assumptions do apply (but references are provided for the curious reader). Some items, such as the encoder and decoder, are only described in words. The chapter's first section on fundamentals may be enough of an introduction to communications theory for those who only want to know what "*I*" and "*Q*" signal components are, how spectrum and filtering are defined, and how bandwidth is defined. The reader is assumed to know what the Fourier transform is.

The topics covered in this chapter are the following:

- *Section 10.2*: Fundamentals of digital communications theory: signal representation, spectrum, filtering, and the end-to-end communications system
- *Section 10.3*: Modulating transmitter, including coding

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

- *Section 10.4*: Filtering
- *Section 10.5*: Demodulating receiver, including decoding
- *Section 10.6*: SNR, E_s/N_0, and E_b/N_0
- *Section 10.7*: Summary of signal distortion sources in end-to-end communications system
- *Appendix 10.A*: Sketch of proof that modulation pulse and signal spectrum are related.

10.2 COMMUNICATIONS THEORY FUNDAMENTALS

10.2.1 Signal Representation

10.2.1.1 RF Signal Representation An RF or **bandpass signal** is a (real-valued) signal for which there is a frequency band about 0 Hz in which it has no power. The signal's **carrier frequency** can be any of the (positive) frequencies in the signal's frequency band, chosen for convenience, but in digital communications the carrier frequency is the center frequency. The RF signal $x(t)$ with carrier frequency f_c in Hz can be represented as follows:

$$\begin{aligned} x(t) &= \sqrt{2}\,\mathrm{Re}\left\{\sqrt{P(t)}e^{j\theta(t)}e^{j2\pi f_c t}\right\} = \sqrt{2}\sqrt{P(t)}\cos[\theta(t)+2\pi f_c t] \\ &= \sqrt{2}\sqrt{P(t)}\cos\theta(t)\cos(2\pi f_c t) - \sqrt{2}\sqrt{P(t)}\sin\theta(t)\sin(2\pi f_c t) \end{aligned}$$

where $\mathrm{Re}(u)$ is defined as the real part of u and f_c is much greater than the rate of change in $\sqrt{P(t)}$ and $\theta(t)$. $P(t)$ is the power of the signal at time t. Any variation in $\sqrt{P(t)}$ with t represents **amplitude modulation (AM)**, and any variation in $\theta(t)$ represents **phase modulation (PM)**. Digital information may be carried in amplitude or phase or both, depending on the type of modulation.

The **rotating phasor** (Couch, 1990) representation of the RF signal is given by the complex function

$$\begin{aligned} &y(t) \text{ for which } x(t) = \sqrt{2}\,\mathrm{Re}\,y(t) \\ &y(t) = \sqrt{P(t)}e^{j\theta(t)}e^{j2\pi f_c t} = \sqrt{P(t)}\cos[\theta(t)+2\pi f_c t] + j\sqrt{P(t)}\sin[\theta(t)+2\pi f_c t] \end{aligned}$$

(Actually, in Couch (1990) the rotating phasor is defined as $\sqrt{2}$ times this, but the present author likes the rotating phasor to have the same power as the signal.) $P(t)$ is the rotating phasor's power at time t, and $\theta(t)$ as a function of time is a phase that jumps back and forth once in a while but averages to zero. Figure 10.1a shows the rotating phasor rotating around at frequency f_c Hz. Figure 10.1b shows an example of $\theta(t)+2\pi f_c t$ as a function of time.

10.2.1.2 RF Signal Equivalent Baseband Representation The fact that the rotating phasor is rotating is not often of interest when we want to analyze the signal. So let us remove the average rotation from our thoughts. Now we have in

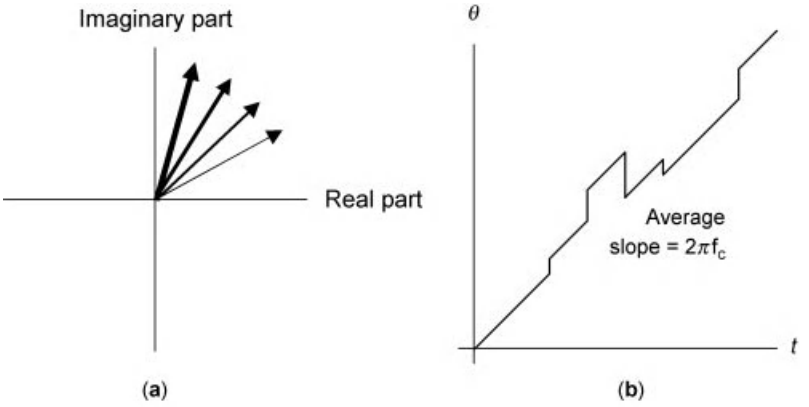


FIGURE 10.1 (a) Rotating phasor represented by arrow; previous positions are represented by thinner arrows and (b) example of phase θ of rotating phasor versus time.

mind the **phasor** (at baseband), given at time t by the complex function

$$z(t) = \sqrt{P(t)}e^{j\theta(t)} = \sqrt{P(t)}\cos\theta(t) + j\sqrt{P(t)}\sin\theta(t)$$

(Most authors define the phasor as $\sqrt{2}$ times this, but the present author likes the phasor to have the same power as the RF signal.) The phasor as a function of t is the **(complex) baseband equivalent** of the RF signal. (The baseband equivalent signal multiplied by $\sqrt{2}$ is called the "complex envelope" of the RF signal, but the author finds this terminology undescriptive.) Figure 10.2a shows the baseband phase versus time corresponding to the RF phase versus time shown in Figure 10.1b. In the interesting case where the phase jumps are multiples of $\pi/2$ and we take the phase modulo 2π, we obtain Figure 10.1b.

The real part of the phasor is called the **in-phase** or ***I* component** and the imaginary part the **quadrature-phase** or ***Q* component**. Referring to the equation we note that

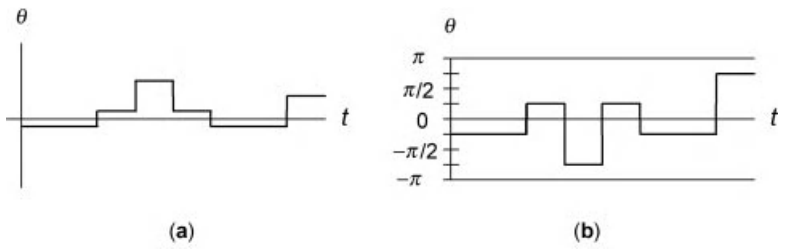


FIGURE 10.2 Same example, phase θ of signal phasor versus time (a) unscaled and (b) scaled so phase jumps are multiples of $\pi/2$, then phase taken modulo 2π.

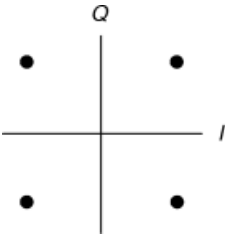


FIGURE 10.3 Same example of signal phasors, shown as dots, when amplitude is constant.

$$I \text{ component} = \sqrt{P(t)}\cos\theta(t) = \frac{1}{\sqrt{2}} \text{ coefficient on } \cos(2\pi f_c t) \text{ term in RF signal}$$

$$Q \text{ component} = \sqrt{P(t)}\sin\theta(t) = \frac{1}{\sqrt{2}} \text{ coefficient on } (-\sin(2\pi f_c t)) \text{ term in RF signal}$$

When, in addition, the amplitude stays constant, the signal phasors take on only four possible values, shown in Figure 10.3. The I and Q components of the signal versus time are shown in Figure 10.4.

A **native baseband signal** is something different from the baseband equivalent of an RF signal. A native baseband signal is real-valued and has all its power in the neighborhood of 0 Hz.

10.2.2 Spectrum Fundamentals

The signal spectrum is a signal's **power spectral density** (**psd**) function versus frequency, which has units of power per Hz or some logarithmic variation of this. The psd is real-valued. The integral over any frequency band tells us how much power of the signal is within that band. The total integral of the psd is the signal power P. We write $\mathcal{S}_u(f)$ for the psd of any signal $u(t)$.

The psd of an RF signal has several representations. There are two RF representations, shown by example in Figure 10.5a and b. The preferred one, the **(one-sided) RF spectrum**, is in Figure 10.5b. Figure 10.5c shows the psd of the baseband equivalent signal. It is the **(two-sided) equivalent-baseband spectrum**. The example

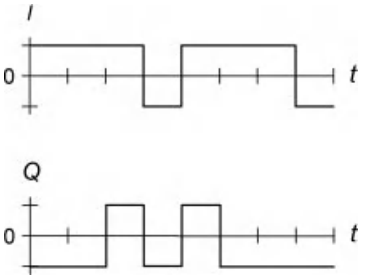


FIGURE 10.4 Same example, I and Q components of same baseband signal versus time.

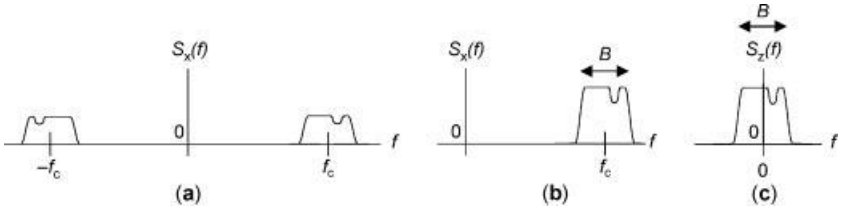


FIGURE 10.5 RF signal's spectrum: (a) two-sided RF representation (not usual); (b) one-sided RF (preferred); and (c) two-sided equivalent-baseband.

spectrum has a notch for demonstration purposes. All the psds integrate to P. In the preferred RF representation and in the equivalent-baseband representation the bandwidths B are equal (while for the unpreferred RF representation the bandwidth is double, since the psd is half as great).

The psd of a native baseband signal has two representations, shown in Figure 10.6, where the preferred one, the **(one-sided) spectrum**, is shown in Figure 10.6b. Both psds integrate to P. In the preferred representation the bandwidth B is one-sided (for the unpreferred representation the bandwidth is double, since the psd is half as great).

The author has been in conversations about "the" bandwidth of a spectrum and later realized that the other person and she were speaking of different things, by a factor of two. Companies tend to settle on a standard, but if you are new, ask which is meant.

Aside from the issue of whether the bandwidth is two-sided or one-sided, there are various types of spectrum bandwidth, for different purposes. The most common for satellite communications are probably the **3-dB bandwidth** and the **mainlobe bandwidth**, all illustrated for the equivalent baseband spectrum of an RF signal in Figure 10.7. The 3-dB bandwidth is the width of that part of the spectrum over which the spectrum is greater than -3 dB or 0.5 of the peak value. The mainlobe bandwidth is the width of the spectrum's mainlobe (most actual spectra have a mainlobe and sidelobes), which typically contains almost all the signal power.

We must also consider the spectrum of noise. **White Gaussian noise** is a theoretical construct, noise of a constant-level psd over all frequencies. As a function of

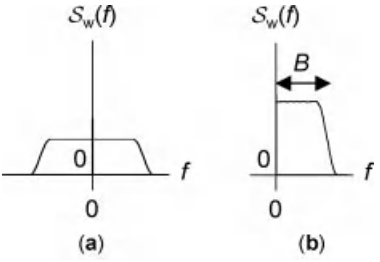


FIGURE 10.6 Native baseband signal's spectrum (a) two-sided representation (not usual) and (b) one-sided (preferred).

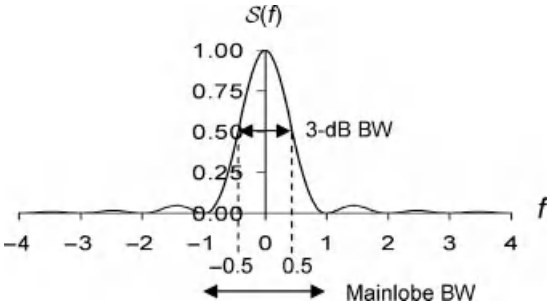


FIGURE 10.7 Definitions of two types of spectrum bandwidth, for equivalent-baseband spectrum of RF signal.

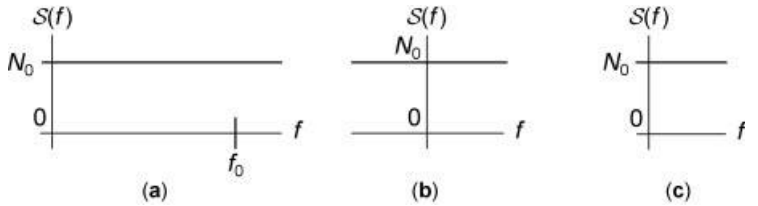


FIGURE 10.8 Noise psd in preferred representation: (a) RF, (b) baseband equivalent of RF, and (c) native baseband.

time it is a random process, meaning a set of random variables indexed by time (see, e.g., (Couch, 1990)). When filtered by a bandpass filter, its probability density function is Gaussian (Appendix 8.A.5). Figure 10.8 shows the noise psd for all three kinds of spectrum, each in its preferred representation; the level is N_0 in all cases.

10.2.3 Filtering Fundamentals

10.2.3.1 Filter Representation To start on filters we need to know the (**Dirac) delta function**, written $\delta(t)$ in the time domain. It is a useful theoretical construct. The function is zero everywhere except where the argument equals zero, at which it is infinite. The integral over its full argument range equals 1. The function in the time domain is drawn in Figure 10.9, where the arrow at zero argument reaching to the level of 1 means an infinite value with the integral of 1. The delta

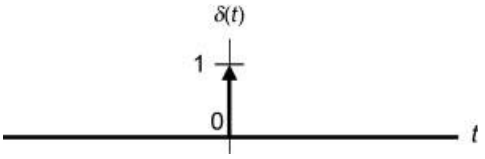


FIGURE 10.9 Dirac delta function in time domain.

function can alternatively be defined as the only function having the following property (Couch, 1990):

$$\int_{-\infty}^{\infty} \delta(\tau)u(t-\tau)d\tau = u(t) \quad \text{for all function } u(\cdot) \text{ continuous at } t$$

Suppose that $h(t)$ is the **impulse response** function of a filter, that is, the output of the filter when the input is the delta function. The filter may be at RF or baseband. When $u(t)$ is a signal input to the filter, the signal $v(t)$ out of the filter is given for all t by

$$v(t) = (h \circ u)(t) \triangleq \int_{-\infty}^{\infty} h(\tau)u(t-\tau)d\tau = \int_{-\infty}^{\infty} u(\tau)h(t-\tau)d\tau = (u \circ h)(t)$$

where $\circ$ is **convolution**. The effect of the filter on the signal is to smear the signal in time. Filtering is a linear operation on the input signal, so it induces **linear distortion**.

The **transfer function** $H(f)$ of the filter is the Fourier transform of $h(t)$, where f is frequency. The Fourier transform $V(f)$ of $v(t)$ is given for all f by $H(f)U(f)$ and the spectrum $\mathcal{S}_v(f)$ of $v(t)$ is given by $|H(f)|^2\mathcal{S}_u(f)$.

A filter developer characterizes a filter $H(f)$ as $A(f)e^{j\varphi(f)}$ for $A(f) \geq 0$ and $\varphi(f)$ real-valued. $A(f)$ and $\varphi(f)$ are, respectively, the **amplitude** and **phase responses** of the filter's transfer function. The name "amplitude response" is a shortcut: $A(f)$ is not actually an amplitude versus frequency but a multiplier on the amplitude of the signal's Fourier transform. Most of the time such a multiplier is stated as the filter's **gain response** $G(f)$ in dB, equal to $20\log A(f)$, where log is defined as $\log_{10}$. The phase response is normally given in degrees.

Most of the time we think of an RF filter as being centered around some RF frequency f_0. Figure 10.10 shows two representations of an RF transfer function, (a) the two-sided RF version and (b) the one-sided RF version, which is preferred. The heights of both representations are the same, unlike for the two RF spectrum representations. When we analyze the effect of a filter on a signal, it is easier to do this in terms of their baseband equivalent representations. The **baseband equivalent transfer function** of the RF filter is shown in Figure 10.10c. It is constructed by sliding down the positive-frequency part of the RF transfer function to center it

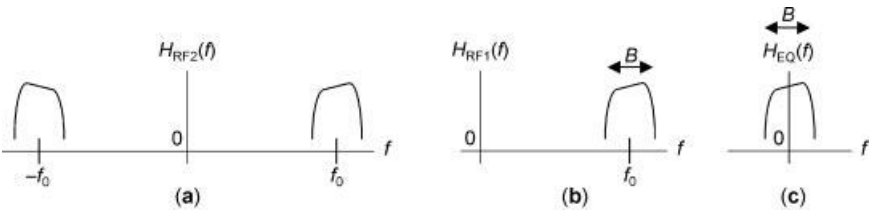


FIGURE 10.10 RF filter's transfer function (a) two-sided RF representation (not usual), (b) one-sided RF (preferred), and (c) baseband equivalent.

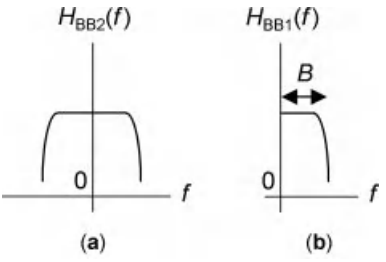


FIGURE 10.11 Native baseband filter's transfer function (a) two-sided representation (not usual) and (b) one-sided (preferred).

about 0 Hz instead of f_0. The baseband equivalent impulse response is not necessarily real-valued. In the preferred RF representation and the baseband equivalent representation the bandwidths B are equal (while for the unpreferred RF representation the bandwidth is double).

Things are different for a native baseband filter, such as the ALC transfer function. Figure 10.11 shows two representations of such a transfer function, (a) the two-sided version and (b) the one-sided version, the preferred one. The impulse response is real-valued. In the preferred representation the bandwidth B is one-sided (for the unpreferred representation the bandwidth is double).

The author has been in conversations about "the" bandwidth of a filter and later realized that the other person and she were speaking of different things, by a factor of two. Usually the other person was inexperienced but even so performing a communications simulation (see Section 13.3.1 for the pitfalls of simulation).

10.2.3.2 Types of Filter Bandwidth Aside from the issue of whether the filter bandwidth is two-sided or one-sided, there are various types of filter bandwidth, for different purposes. The most common for satellite communications are probably the **3-dB bandwidth** and the **(equivalent) noise bandwidth**, illustrated for the baseband equivalent transfer function of an RF signal in Figure 10.12. The 3-dB bandwidth is the maximum bandwidth between two frequencies where the filter's

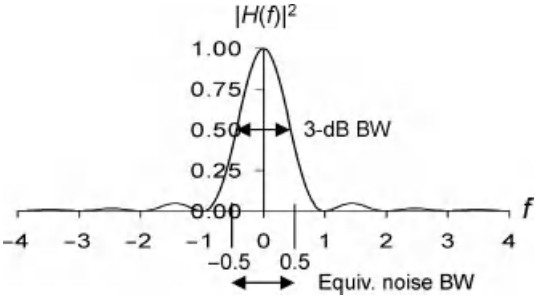


FIGURE 10.12 Definitions of two types of filter bandwidth, for equivalent baseband transfer function of RF filter.

gain response falls no more than 3 dB down from its maximum value inside the band (Couch, 1990). This definition is useful for describing roughly the bandwidth over which the signal is passed, outside of which the signal is mostly not passed. The noise bandwidth of the equivalent baseband transfer function of an RF filter is defined as follows:

$$\text{Noise BW} = \frac{1}{H(0)^2}\int_{-\infty}^{\infty} |H(f)|^2 df$$

This definition is useful in the formula for noise power passed by a filter when the noise psd is flat and wider than the filter. Other perfectly good bandwidth definitions, for other purposes, are given in Couch (1990).

The most common RF filters in the payload are bandpass, as we have been showing in the figures so far. However, there are other kinds of filters: notch, low-pass, and high-pass. These three kinds do not have a bandwidth. A notch filter is described by how many dB of rejection it has over a particular (narrow) frequency band. Low-pass and high-pass filters are described by how much rejection they have at frequencies above and below the signal band, respectively.

10.2.4 End-to-End Communications System Fundamentals

Figure 10.13 gives a simplified diagram of the end-to-end satellite communications system, from input bit stream to output bit stream. The diagram shows the baseband equivalent representation of the RF system, for the sake of simplicity and to match the way we describe the system in the rest of this chapter. So frequency converters and antennas are missing. In terms of communications theory, which calls everything except the modulator and demodulator the "channel" (in a different meaning from in this book), the system is a complicated "channel," primarily because of the nonlinear element in the middle (the HPA) and the **additive white Gaussian noise (AWGN)** at two points instead of the usual one. In the actual system the additive noise comes primarily from the front-end LNA of the payload and ground receiver, respectively.

The basic idea of a communications system is that the elements of the transmitter are mirrored in the receiver by reverse elements. In addition, the receiver has an element (detection filter) to reduce the effect of the added noise and sometimes an

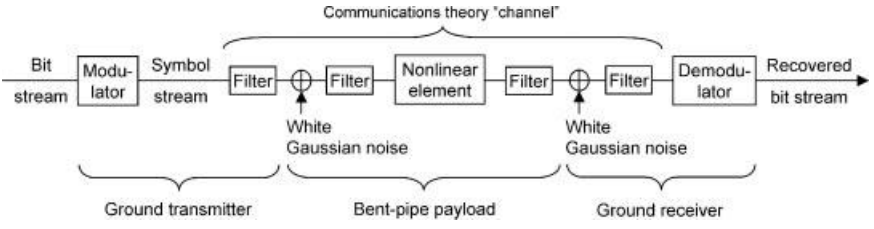


FIGURE 10.13 Simplified drawing of satellite communications system, equivalent-baseband, for bent-pipe payload.

element (equalizer, not described in this book) to reduce the effect of the filtering between transmitter and receiver. The next three sections of this chapter will go into detail about the following:

- *Section 10.3*: Ground transmitter, called the "modulating transmitter" to differentiate it from a bent-pipe payload's transmitter. A processing payload would also have a modulating transmitter.
- *Section 10.4*: Filtering.
- *Section 10.5*: Ground receiver, called the "demodulating receiver" to differentiate it from a bent-pipe payload's receiver. A processing payload would also have a demodulating receiver.
- *Section 10.6*: Signal-to-noise ratio, E_s/N_0, and E_b/N_0 anywhere in the signal path and specifically at input to symbol decision.

10.3 MODULATING TRANSMITTER

10.3.1 Architecture

The input to the transmitter is a stream of data bits arriving at a constant rate. Data bits are also called "information bits." The baseband part of the modulating transmitter is depicted in Figure 10.14. This particular depiction is not the only one possible but is good for exposition purposes. The output of the baseband part of the transmitter is a complex-valued continuous function.

In the next part of the transmitter the IF carrier gets modulated, as drawn in Figure 10.15.

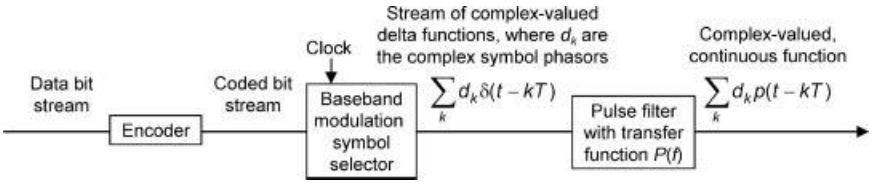


FIGURE 10.14 Baseband part of modulating transmitter.

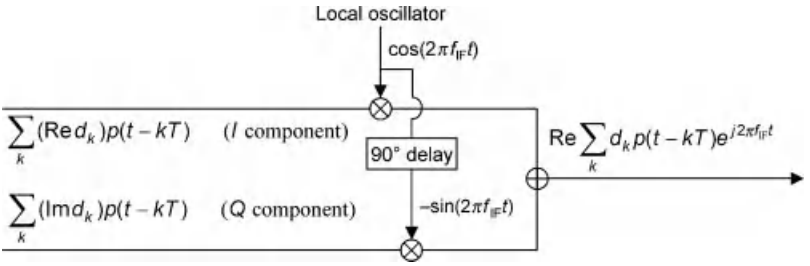


FIGURE 10.15 Modulating the IF carrier.

The last part of the transmitter is not depicted but consists of the upconverter to RF, filters, a HPA (assumed to be operating in its linear amplification region), perhaps a further filter, and the transmitting antenna. This part of the transmitter is not further discussed in this chapter.

In the rest of this section, we focus on the baseband part of the transmitter.

10.3.2 Encoder

An **encoder** is used on most satellite uplinks. It adds redundancy to the data bit stream. The big advantage of coding the data bit stream and decoding is that to achieve the necessary **bit error rate (BER)** (Section 10.5.4), the transmitter power and/or the receiver sensitivity can be dropped by usually several dB altogether. This type of encoder is a "channel encoder" since it protects against the errors introduced by the communications-theory "channel." Another name for it is "error-control coding." In contrast, a "source encoder" has the purpose of reducing redundancy in the bit stream to decrease the transmitted bit rate (Benedetto et al., 1987). If there is a source encoder in the system, it would be outside of the diagram of Figure 10.14 and its output would be the data bit stream input in the diagram. The word "code" can cause confusion because it is used for many different things.

An encoder adds redundancy in such a way that each coded (output) bit is a combination of many input bits and perhaps already coded bits. The idea is that if the receiver "detects" or guesses some coded bits wrongly, the decoder can use all the relationships to still reconstruct the original data bit stream correctly. A code is described partially by its **code rate** and its type. The code rate is a ratio $n/m < 1$ of integers n and m with the meaning that the number of coded bits in a time period is m/n times the number of data bits. A code with a rate close to 1 can achieve the same performance as a lower-rate code when the higher-rate encoder does more processing (and similarly the decoder does more processing). Now, the data-processing part of a receiver (outside our scope) requires a certain BER to properly function.

There are three types of channel encoder: the conventional **block encoder** and **convolutional encoder** and the newer turbo encoder. The conventional encoders are described in Couch (1990) and more fully in Lin and Costello Jr. (1983), while turbo encoders are described in Proakis and Salehi (2008).

The most common types of block coding are RS (Reed–Solomon) and BCH (Bose, Chaudhuri, and Hocquenghem). Both take a fixed number of data bits and compute a larger block of coded bits.

Convolutional coding on the other hand is a continuous process. It may first divide the input bit stream into two or three parallel streams, say k of them, but k can also equal 1. When a new bit is shifted into each of the k input streams, $i > k$ linear combinations of new and old bits are formed and output. Such a code has rate k/i. If some bits are then removed in a regular pattern, that is, the code is **punctured**, then the rate becomes higher than k/i.

Block and convolutional codes are sometimes **concatenated**, in which one code is applied on top of another in order to achieve even more of a drop in required

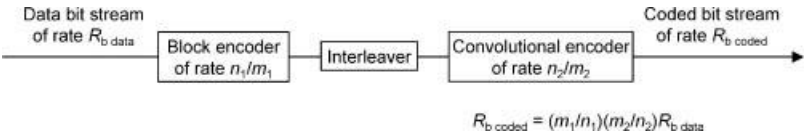


FIGURE 10.16 Concatenated encoder with interleaving in between.

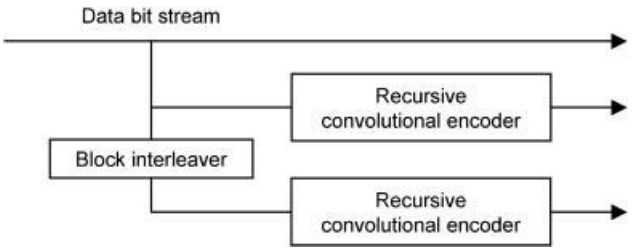


FIGURE 10.17 Turbo encoder (after Proakis and Salehi (2008)).

transmitter power and/or receiver sensitivity for the same BER. The process is shown in Figure 10.16. Usually the first code applied to the data bit stream is a block code and the second is a convolutional code, and there is an **interleaver** in between. The interleaver reorders the bits. It is easier to describe the purpose of concatenated coding in terms of the decoding. The first decoder, for the convolutional code, corrects many of the errors. When it makes a mistake its output is a bit stream with a run of errors, that is, correlated. The second decoder, for the block code, needs as input a bit stream which has uncorrelated errors in it. The deinterleaver breaks up and widely separates the errors, making them uncorrelated as far as the block decoder is concerned. Interleavers come in two types, block and convolutional.

Turbo coding is also a concatenated coding scheme, depicted in Figure 10.17, which uses two convolutional encoders and a block interleaver. Each convolutional encoder is recursive, which means that its output bits are not just a combination of current and earlier input bits but also of earlier output bits (Proakis and Salehi, 2008). Turbo codes perform better than the others but are more complicated to decode.

In general, the coding techniques used in satellite communications are not as advanced as the state-of-the art techniques in terrestrial systems. Two of the reasons are that technology for the state-of-the art techniques would not yet exist at the very high rates typical of satellite communications, and some ground transmitters and receivers are legacy hardware or inexpensive consumer equipment.

10.3.3 Baseband Modulator

Only a few of the large variety of digital modulation schemes are used in satellite communications. The main reason is that satellite bus power is limited, so the HPAs must be operated in a power-efficient manner. When channel sharing is by TDM (Section 1.2.1), usually the HPA cannot be far backed off, if at all; this means that

the HPA operates in its nonlinear region. A useful modulation scheme then has constant-amplitude phasors. (If either a highly spectrum-efficient modulation is required or channel sharing is by FDM then the HPA cannot run near saturation and it is optimized for the well-backed-off region.) Bandwidth is also a limited commodity on the payload, so usually the modulation has to be at least moderately bandwidth-efficient. Ease of implementation on the ground may also be a consideration.

In this subsection, the complex symbol phasors d_k (see Figure 10.14) of each modulation scheme are given. It is a good question why this is even called "modulation," albeit "baseband." It is because the phasors represent the selection of the phase and in some cases amplitude for the carrier frequency.

The allocation of bits to particular phasors in some modulation schemes, called "Gray coding," is discussed after the modulation schemes.

10.3.3.1 Modulation Schemes The modulations described in this subsection are the ones commonly used over satellite (Maral and Bousquet, 2002; Elbert, 2004) plus an upcoming one:

- QPSK
- OQPSK
- BPSK
- 8PSK
- 16QAM
- 16APSK
- MSK.

(Apparently $\pi/4$-QPSK (Maral and Bousquet, 2002) was proposed for use over satellite but has not found use.)

Quaternary phase-shift keying (QPSK) and its slight variation **offset QPSK (OQPSK)** may be the modulation schemes most commonly used in satellite communications (Elbert, 2004). The reasons are that they are power-efficient and, ideally, anyway, of constant amplitude so suffer little degradation from the HPA's nonlinearity. They are also moderately bandwidth-efficient and simple to implement. Their four phasors are depicted in the leftmost drawing of Figure 10.18. The

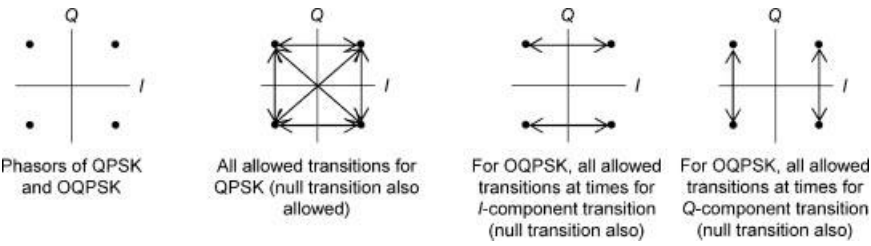


FIGURE 10.18 QPSK and OQPSK.

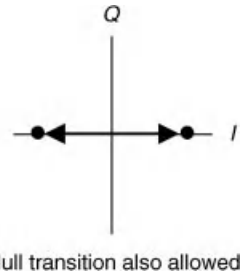


FIGURE 10.19 BPSK.

next drawing to the right depicts, for QPSK, all allowed transitions between phasors. In OQPSK or staggered QPSK (SQPSK), the *I* and *Q* phasor components transition sequentially instead of at the same time, as depicted in the rightmost two drawings of the figure. In a system with significant filtering before the HPA, QPSK would show a larger amplitude variation than OQPSK would, which means greater distortion by the HPA, so OQPSK would be preferred.

Binary phase-shift keying (BPSK) is "half" of QPSK in that only one component, call it "*I*," of the phasor is created, as shown in Figure 10.19. Because BPSK is less spectrally efficient than QPSK (i.e., for the same bandwidth you only get half the bit rate), it is used less often than QPSK. It does, however, require only half the transmit power of QPSK for the same BER.

Another PSK modulation is **8-ary phase-shift keying (8PSK)**. Instead of only four phasors as in QPSK, there are eight, evenly spaced around the circle as illustrated in Figure 10.20. While this modulation scheme is spectrally efficient, it is not power-efficient. From any phasor, any of the eight can be reached next.

A step up in complexity from PSK is **16-ary quadrature-amplitude (16QAM)** modulation, since it has not just phase modulation but also amplitude modulation. The phasors are arranged in a regularly spaced 4×4 matrix as shown in Figure 10.21. From any phasor, any of the 16 can be reached next. 16QAM is spectrally efficient and power-efficient but is only usable when the HPA is backed off into a near-linear region.

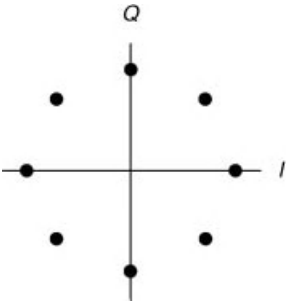


FIGURE 10.20 8PSK phasors.

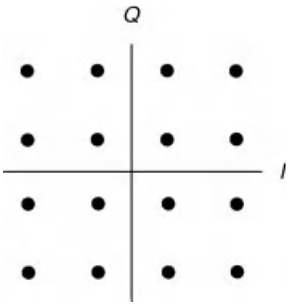


FIGURE 10.21 16QAM phasors.

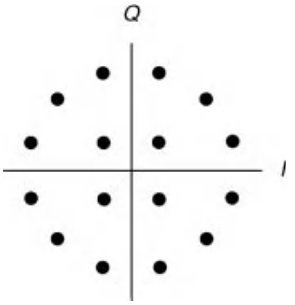


FIGURE 10.22 16APSK phasors.

Similar to 16QAM but more power-efficient for single-carrier HPA operation is **16-ary amplitude-and-phase-shift keying (16APSK)** modulation, shown in Figure 10.22. The phasors are in two rings, where often the power of the outer ring is set so as to saturate the HPA. From any phasor, any of the 16 can be reached next. The modulation is also known as "16APK," "(12,4) PSK," and "12/4 QAM." It was introduced in a standard for broadband satellite applications in the mid-2000s (latest version (ETSI, 2009)) and has been much studied before and since.

Minimum-shift keying (**MSK**) has been used in recent years to take better advantage of the limited power of SSPAs (Elbert, 2004). With MSK the phasor is always moving at a constant rate along a circle either clockwise or counter-clockwise, as illustrated in Figure 10.23. In the literature, it is usually compared to

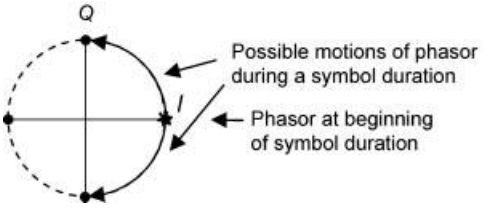


FIGURE 10.23 MSK smooth transitions allowed during symbol duration, from one of four possible starting phasors.

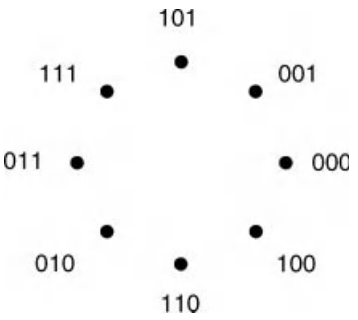


FIGURE 10.24 Example of Gray coding.

OQPSK. The mainlobe of the MSK spectrum is 1.5 times as wide as that of OQPSK, so MSK has low bandwidth efficiency.

We must specialize the expression for the equivalent-baseband signal given in Figure 10.14 to the particular modulation scheme. For QPSK, BPSK, 8PSK, 16QAM, and 16APSK, the complex baseband signal is as in the figure, namely $\sum_k d_k \delta(t - kT)$, where d_k is the kth complex phasor. For OQPSK and MSK, where I and Q bits are used sequentially, it is more accurately written as $\sum_k \left[d_{kI}\, p(t - kT) + j d_{kQ} p(t - (k + \frac{1}{2})T) \right]$, where d_{kI} and d_{kQ} are the kth I bit and Q bit, respectively.

10.3.3.2 Gray Coding Some modulation schemes need a rule for how to assign coded bits from the single stream to phasors. For BPSK, OQPSK, and MSK no rule is needed, as only one bit at a time is used from the bit stream. However, QPSK, 8PSK, 16QAM, and 16APSK use more than one bit at a time to determine the next signal phasor, so they must break up the bit stream into small groups. The number of bits in a group is two for QPSK, three for 8PSK, and four for 16QAM and 16APSK. The commonly used assignment method, **Gray coding**, leads to a smaller BER than any other assignment does. Gray coding is based on the fact that the most likely kind of error that the receiver will make in deciding what phasor was transmitted is to mistake a corrupted phasor for one of the transmitted phasor's nearest neighbors. This is because smaller distortion is more likely than larger distortion. In Gray coding, the bit groups assigned to nearest-neighbor phasors differ in only one bit, so the most likely kind of decision error will correspond to only one bit error. Figure 10.24 gives an example of Gray coding for 8PSK. There is more than one way to Gray code.

10.3.4 Pulse Filter and Signal Spectrum

10.3.4.1 Introduction The modulation phasors have been depicted in Figure 10.14 as being coefficients on a stream of delta functions, and now we need to turn this stream into a (baseband-equivalent) signal. Figure 10.14 shows that this

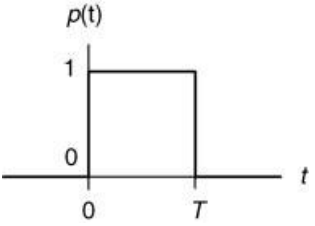


FIGURE 10.25 Rectangular pulse $p(t)$ of duration T.

complex-valued stream gets filtered by a filter representing the desired **pulse**. We treat only real-valued pulses. The choice of pulse depends on ease of implementation, how much bandwidth is available, how much filtering the signal will see, and how far backed off the HPA will be. For all the modulations except MSK the choice is open; for MSK it is made.

It helps in this introductory discussion to have a pulse in mind, so we consider the **rectangular pulse**, shown in Figure 10.25 (but more on this pulse in the next subsection). The pulse equals 1 from time zero to time T and equals zero elsewhere. With this pulse the signal jumps instantaneously to its next phasor and stays exactly at that phasor until T later.

We define the **(modulation-)symbol duration** as T and the **(modulation-)symbol rate R_s** as $1/T$. The **bit rate R_b** equals R_s for BPSK; $2R_s$ for QPSK, OQPSK, and MSK; $3R_s$ for 8PSK; and $4R_s$ for 16QAM and 16QPSK. The definition of symbol duration will hold even when the pulse is nonzero over a longer time than T; for each modulation scheme T will be shown.

We assume that the coded bits are independent of each other and that each is equally likely to be -1 or 1. Then there exists the following simple relationship, for all the modulation schemes, between the signal spectrum and the pulse transform (shown in Appendix 10.A):

$$\mathcal{S}_z(f) = E\left[|d_k|^2\right]|P(f)|^2 \text{ where } E\left[|d_k|^2\right] \text{ is the average power of the complex phasors}$$
$$\text{and } P(f) = [\mathcal{F}p](t), \text{ with } \mathcal{F} \text{ the Fourier transform operator}$$

We will lapse sometimes and refer to $|P(f)|^2$ as the **pulse spectrum**.

10.3.4.2 Rectangular Pulse The rectangular pulse is a useful fiction. It is not realizable because its jumps at time zero and T are infinitely sharp, for which to create it needs frequencies out to infinity. To be able to plot $P(f)$ we need it to be real-valued, so we shift the pulse back in time to make it symmetric about zero time as shown in Figure 10.26a. (We do this with all the pulses, making them all noncausal (Couch, 1990), but we do not let this bother us, as the resulting explanation simplifications are worth it.) Then the Fourier transform $P(t)$ of the pulse is the $\sin(x)/x$ function as shown in Figure 10.26b and the spectrum is the square of this, as shown in Figure 10.27.

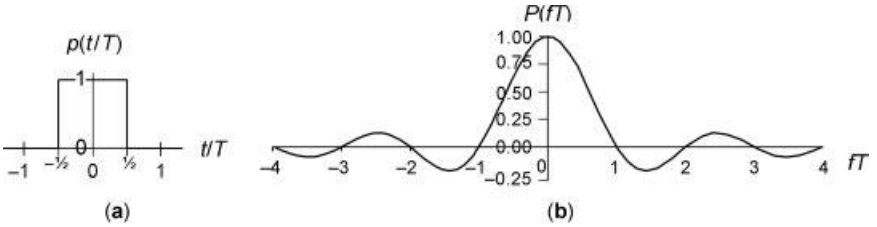


FIGURE 10.26 Rectangular pulse $p(t)$, symmetric about zero time, and its Fourier transform $P(f)$.

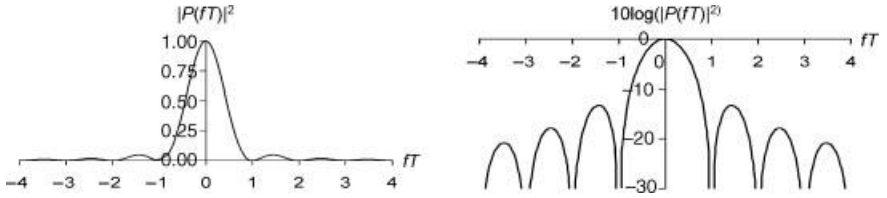


FIGURE 10.27 $|P(f)|^2$ for rectangular pulse, shown both not in dB and in dB.

One way to limit the extent of the pulse spectrum is to chop it off outside of some band. Conceptually we are multiplying the spectrum by another abstraction, the theoretical **brickwall filter** shown in Figure 10.28.

Let us see what this bandwidth chopping does to the pulse. Figure 10.29 shows the resultant pulse for when all but the mainlobe is chopped off. The pulse is nearly

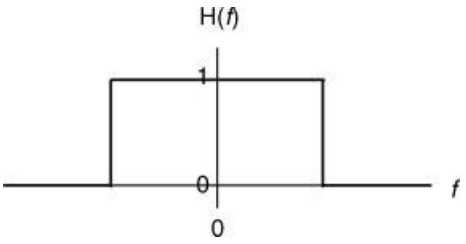


FIGURE 10.28 Transfer function $H(f)$ of theoretical brickwall filter at baseband.

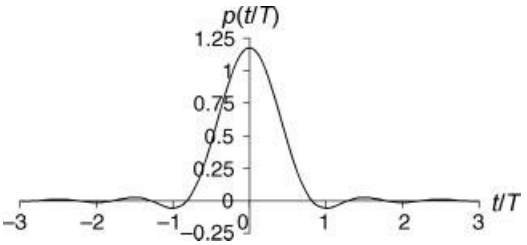


FIGURE 10.29 Bandlimited rectangular pulse $p(t)$, all but spectrum mainlobe filtered out.

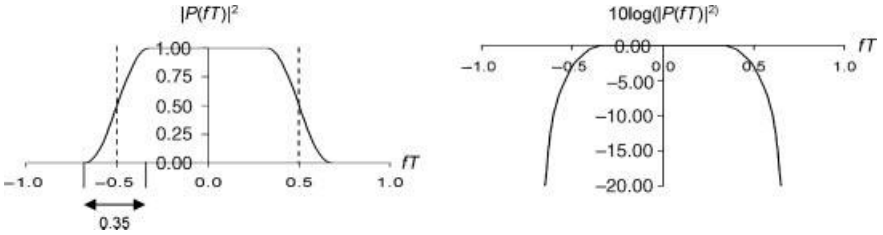


FIGURE 10.30 Pulse spectrum of raised-cosine shape with $\alpha = 0.35$.

triangular, which may be surprising since the mainlobe contains 90% of the signal power. The pulse is near zero outside the triangular part. The result of the bandlimiting on the signal is that the signal phasor no longer just sits at one of the ideal spots as shown in Section 10.3.3.1 and then instantaneously jumps to another ideal spot. Instead, most of the time it is simply near one or another of the ideal spots but it is always moving.

10.3.4.3 Root Raised-Cosine Pulse A more commonly used approach is to start with a good spectrum and use the resultant pulse. This is (somewhat misleadingly) called "pulse-shaping." The most common solution for satellite communications is the **raised-cosine (RC) spectrum**, whose square root is the **root raised-cosine (RRC)** pulse transform. A case of the RC spectrum is shown in Figure 10.30 for one of the commonly used values of the **roll-off factor α** of the raised cosine, namely 0.35. The factor α is defined so that $\alpha = 0$ gives the square spectrum and $\alpha = 1$ gives a purely raised-cosine shape, that is, no flat area. Figure 10.31 shows the pulse for $\alpha = 0.35$ and for the other commonly used value of 0.2. A small α like 0.2 gives a very narrow spectrum but more signal-amplitude variation than the larger α does, leading to more nonlinear distortion by the satellite's HPA.

10.3.4.4 MSK Pulse A way to see what the MSK pulse is, is to think of the signal starting at one of the phasors on the I axis, which means it is time to act on a Q bit. If the next Q bit is $+1$ then the Q-component over the duration T will be a positive half-cycle of sine; if it is -1 the Q-component will be a negative half-cycle of sine. The pulse, shifted back by $T/2$ so its transform is real-valued, is illustrated in Figure 10.32, and the pulse spectrum is shown in Figure 10.33.

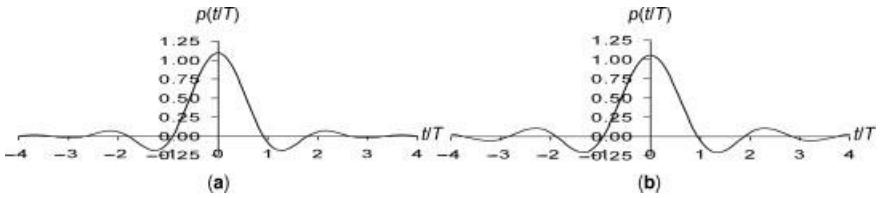


FIGURE 10.31 Pulse with RRC transform with (a) $\alpha = 0.35$ and (b) $\alpha = 0.2$.

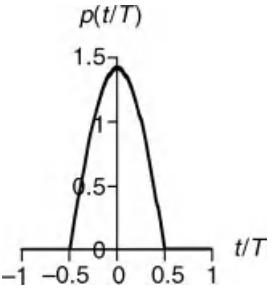


FIGURE 10.32 MSK pulse.

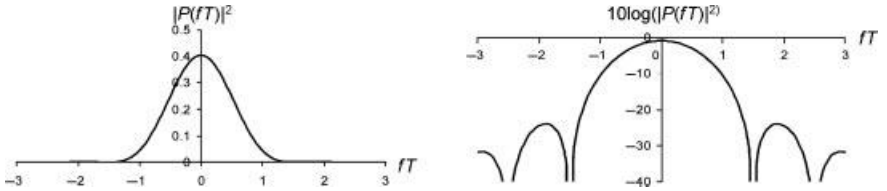


FIGURE 10.33 MSK pulse spectrum.

10.4 FILTERS

Between the modulator and demodulator, the signal will pass through many filters. Every device has some bandwidth limitation, that is, only passes some range of frequencies, so every device performs a filtering operation. The filter discussion below treats the baseband equivalent of RF filters and native baseband filters.

10.4.1 Definitions of Even, Odd, Conjugate Symmetric, and Conjugate Antisymmetric Functions

There are two mathematical categories of filters that are important in the next subsection, so we define them here.

There are the terms **even function** and **odd function**, which apply to real-valued functions. A function $x(f)$ is even when $x(f)$ is symmetric about zero, that is, when $x(-f) = x(f)$. A function is odd when $x(-f) = -x(f)$. Every real function $x(f)$ can be separated into an even part $x_e(f)$ and an odd part $x_o(f)$ by the following method:

$$x(f) = x_e(f) + x_o(f)$$

where

$$x_e(f) = \tfrac{1}{2}[x(f) + x(-f)] \quad \text{and} \quad x_o(f) = \tfrac{1}{2}[x(f) - x(-f)]$$

Figure 10.34 gives an example of this separation.

Similar terms for complex-valued functions are **conjugate symmetric** (or Hermitian symmetric (Proakis and Salehi, 2008)) and **conjugate antisymmetric**. A complex function $H(f)$ is conjugate symmetric when $H(-f) = H^*(f)$ and is

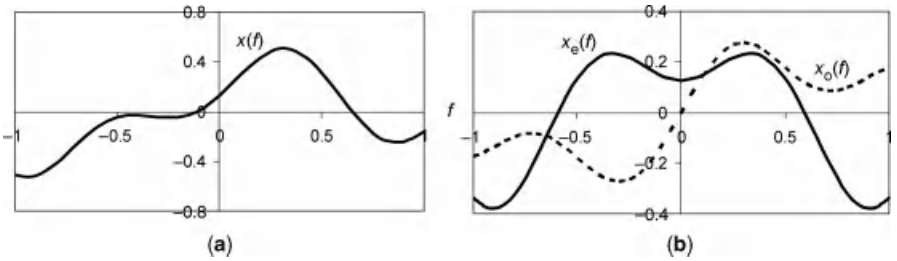


FIGURE 10.34 Example of (a) function, (b) separated into even and odd parts.

conjugate antisymmetric when $H(-f) = -H^*(f)$. Every complex function $H(f)$ can be separated into a conjugate symmetric part $H_{cs}(f)$ and a conjugate antisymmetric part $H_{ca}(f)$ by the following method:

$$H(f) = H_{cs}(f) + H_{ca}(f)$$

where

$$H_{cs}(f) = \tfrac{1}{2}[H(f) + H^*(-f)] \quad \text{and} \quad H_{ca}(f) = \tfrac{1}{2}[H(f) - H^*(-f)]$$

where "*" denotes complex conjugation.

10.4.2 Real and Imaginary Impulse Responses and Intersymbol Interference

The importance of this separation of the transfer function $H(f)$ is the following:

$$H(f) \text{ is conjugate symmetric} \quad \Leftrightarrow h(t) \text{ is real}$$
$$H(f) \text{ is conjugate antisymmetric} \Leftrightarrow h(t) \text{ is imaginary}$$

Part of $H(f)$ creates the real part $h_r(t)$ of the impulse response $h(t)$ and the rest of $H(f)$ creates the imaginary part $h_i(t)$. We divide a signal $u(t)$ into its real part $u_r(t)$ and imaginary part $u_i(t)$. From the formula for convolution, we have

$$\begin{aligned} v(t) = (h \circ u)(t) &= \int_{-\infty}^{\infty} u(\tau)h(t-\tau)d\tau \\ &= \int_{-\infty}^{\infty} [u_r(\tau) + ju_i(\tau)][h_r(t-\tau) + jh_i(t-\tau)]d\tau \\ &= \int_{-\infty}^{\infty} [u_r(\tau) + ju_i(\tau)]h_r(t-\tau) + [-u_i(\tau) + ju_r(\tau)]h_i(t-\tau)d\tau \end{aligned}$$

Recall that a filter "smears" a signal in time. A filter creates **intersymbol interference** (**ISI**) because it smears a signal segment of duration T into nearby such segments. (ISI is not necessarily bad when it exists before the demodulating receiver, unless it increases the signal's amplitude variation and the HPA's distortion. Any residual ISI at detection filter output is bad, however.) A

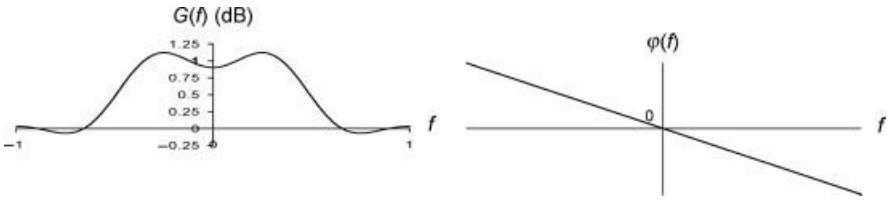


FIGURE 10.35 Examples of filter gain response $G(f)$ and phase response $\varphi(f)$ that cause *I–I* and *Q–Q* crosstalk.

real-valued impulse response creates *I*-to-*I* (i.e., real-to-real) and *Q*-to-*Q* (i.e., imaginary-to-imaginary) ISI. An imaginary-valued impulse response creates *I*-to-*Q* and *Q*-to-*I* ISI, called ***I – Q* crosstalk**.

We want to see how to recognize a real or imaginary impulse response function from the way the filter is characterized by the filter developer, namely by its gain response $G(f)$ and phase response $\varphi(f)$. The natural logarithm of $H(f)$ is ((ln10)/20) $G(f) + j\varphi(f)$. We divide $G(f)$ and $\varphi(f)$ into their even and odd parts, $G(f) = G_e(f) + G_o(f)$ and $\varphi(f) = \varphi_e(f) + \varphi_o(f)$. We now have to assume that ((ln10)/20)$G(f)$, which is about $0.115G(f)$, is small and that $\varphi(f)$, which is in radians, is also small. $\varphi(f)$ is in the interval $(-\pi,\pi]$. By applying the Taylor series three times, we obtain

$$H_{cs}(f) \approx 1 + \frac{\ln 10}{20} G_e(f) + j\varphi_o(f) \quad \text{and} \quad H_{as}(f) \approx \frac{\ln 10}{20} G_o(f) + j\varphi_e(f)$$

This means that for small gain (in dB, multiplied by 0.115) responses and small phase (in rads) responses,

- The even part of the gain response and the odd part of the phase response create *I*-to-*I* and *Q*-to-*Q* ISI.
- The odd part of the gain response and the even part of the phase response create *I–Q* crosstalk.

Figures 10.35 and 10.36 show examples of baseband transfer functions that exhibit these characteristics. The phase response $\varphi(f)$ in Figure 10.35 is a straight line, which corresponds simply to a delay, because the filter's group delay is related to phase by $\tau_g = -d\varphi(f)/(2\pi df)$ (Couch, 1990).

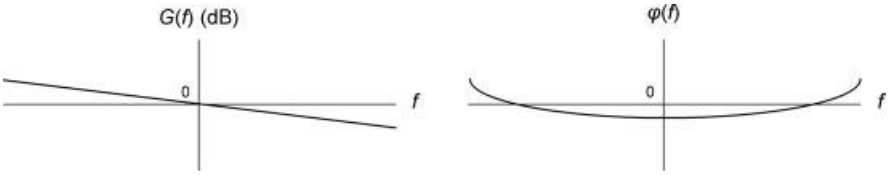


FIGURE 10.36 Examples of $G(f)$ and $\varphi(f)$ that cause *I–Q* crosstalk.

10.5 DEMODULATING RECEIVER

This section is written for a bent-pipe payload, but for a communications system with a processing payload, which has two demodulating receivers, the necessary changes should be apparent.

10.5.1 Architecture

The receiver in a communications system has the job of deciding what the original data bit stream most likely was. The architecture of the demodulating receiver is as follows. The first part (not depicted) consists of a receive filter, a LNA, perhaps more filters, an ALC, and the downconverter to IF. Following that is the demodulation of the distorted and noisy IF signal, as illustrated in Figure 10.37.

In the rest of this section, we focus on carrier recovery and the receiver's baseband part, depicted in Figure 10.38.

10.5.2 Carrier Recovery

Demodulating the received IF signal requires good estimates of the phase, frequency, and perhaps frequency rate of the signal's carrier. **Carrier recovery** or carrier tracking provides these. The element in Figure 10.37 labeled "carrier recovery" is actually only part of the evident carrier-recovery feedback loop.

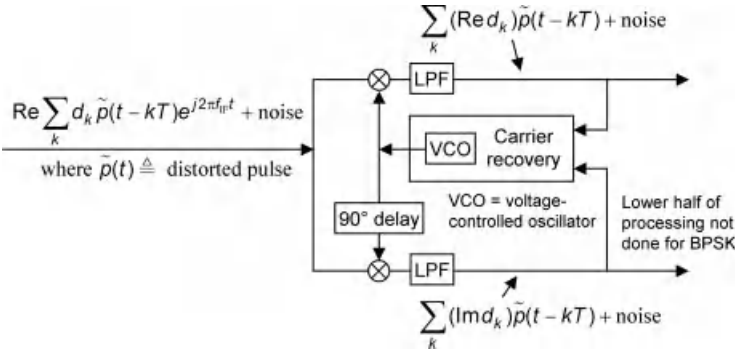


FIGURE 10.37 Demodulating the received IF signal.

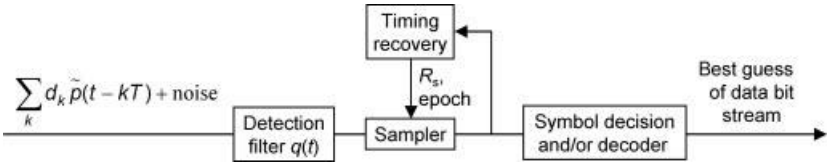


FIGURE 10.38 Baseband part of demodulating receiver.

We have already introduced, in Section 5.5.6.1, the basic phase-locked loop (PLL) that reconstructs a carrier from a noisy input carrier. When the carrier is digitally modulated, there is no residual tone handy to lock to, so it must be created then it can be fed to the PLL. This is discussed in some detail in Proakis and Salehi (2008).

A feedback loop is designed to have a particular **order**, which determines what the loop can track and therefore what the demodulator can remove: a first-order loop can only track a varying phase offset, a second-order loop varying frequency and phase offsets, and a third-order loop varying frequency-rate, frequency, and phase offsets (Gardner, 1979). A second-order loop is sufficient for a GEO satellite, since it can track the small and slowly varying frequency variation from a small orbital inclination. For non-GEO satellites, a third-order loop may be required.

A second-order feedback loop has a native-baseband transfer function (Section 10.2.3.1) with (one-sided) bandwidth B_L. In the typical case of a high-gain second-order loop with damping factor $\zeta = 0.707$, the 3-dB-down point of the closed-loop response is at $B_L/3.33$ (Gardner, 1979). Approximately speaking, the loop tracks phase noise and spurious PM that have a variation rate less than $B_L/3.33$ and cannot track components that have a faster rate, as illustrated in Figure 10.39. Any phase noise at a rate higher than roughly $R_s/2$ will be averaged out by the detection filter, nullifying its effect. The detection filter is, roughly speaking, an integration over the symbol duration $T = 1/R_s$, which has a noise bandwidth at baseband of about $R_s/2$ (one-sided). For more accurate calculations or other ζ, refer to Gardner (1979).

Phase noise has several sources in the ground transmitter, the bent-pipe payload, and the ground receiver, as illustrated in Figure 10.40. There are variations on the illustrated case; for example, the ground transmitter may modulate directly to RF, the payload may also have an upconversion, and an upconversion and a downconversion oscillator may be derived from the same source, which reduces phase noise (Section 5.5.2).

Designing the carrier-tracking loop including picking an appropriate value for B_L is a job for a ground station designer with knowledge of the phase noise

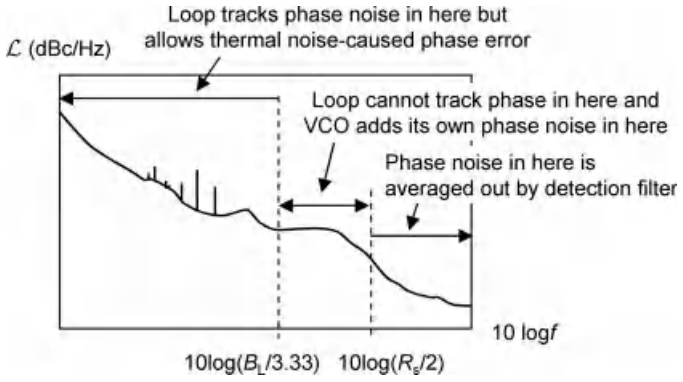


FIGURE 10.39 Approximately what happens to phase noise and phase spurs in various parts of spectrum.

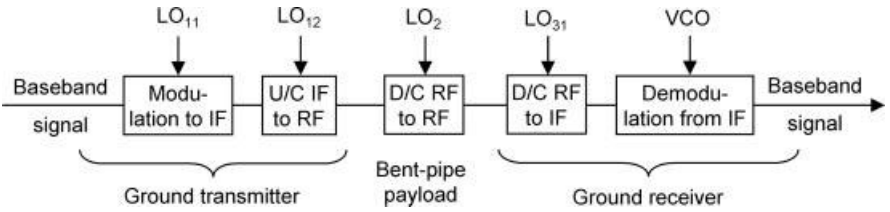


FIGURE 10.40 Phase noise sources in bent-pipe payload communications system.

characteristics of all system oscillators. For a second-order loop, choosing B_L is a tradeoff between tracking as much as possible of the phase noise (large B_L) and minimizing the thermal noise-caused phase error (small B_L).

Phase noise causes a zero-mean, Gaussian-distributed error on the signal's phase, as illustrated in Figure 10.41 for QPSK. While it is unlikely that phase noise itself will cause the wrong decision on what phasor was sent, in combination with thermal noise, signal distortion, and interference it will cause a wrong decision now and then.

10.5.3 Detection Filter and Sampler

The **detection filter** is the filter in the demodulating receiver that cleans up the signal prior to sampling and symbol decision, basically by maximizing SNR. In this subsection, we look at what this filter should be, what the filtered signal looks like, when to sample the output for the following symbol decision element, and what the timing recovery element is.

10.5.3.1 What to Use as Detection Filter Figure 10.42 shows a simplified equivalent-baseband model of the middle part of the end-to-end satellite communications system, from pulse filter $P(f)$ through to detection filter $Q(f)$ (compare to

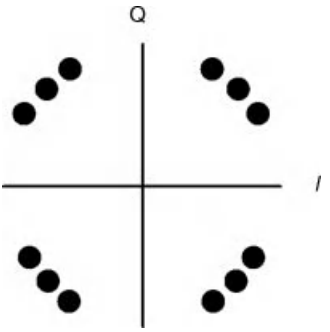


FIGURE 10.41 QPSK phasors with phase noise on them.

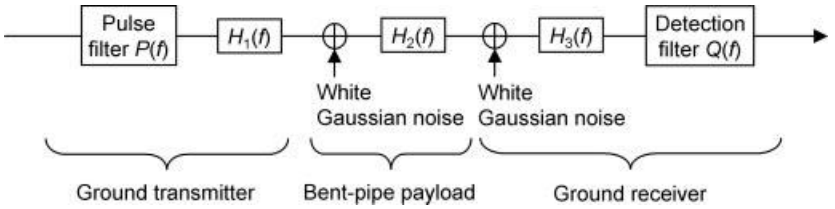


FIGURE 10.42 Middle part of simplified equivalent-baseband model of end-to-end satellite communications system.

Figure 10.13). We ignore the HPA's nonlinearity but its filtering aspect (Section 13.3.6) is incorporated into filter $H_2(f)$. The AWGN enters in two places, in the satellite electronics and in the ground electronics. Any noise-like interference is included in the Gaussian noise.

The question of what the detection filter should be boils down to whether more noise is added on the uplink or the downlink, relative to the signal power. Usually one or the other noise is dominant; it does not often happen that they are nearly equal. If only one link involves consumer equipment, usually that link has the dominant noise. Now, when the only signal impairment is one-site AWGN, the optimal detection filter is the **matched filter (MF)**, which maximizes the SNR into the symbol detector. It is the complex conjugate of the total filtering that the signal has seen before the AWGN times the inverse of the filtering after (see, e.g., Proakis and Salehi (2008) and Couch (1990)). For us, neglecting the nondominant noise source for now, $Q(f)$ would be the following for the two cases of the dominant noise location:

$$Q(f) = \begin{cases} P^*(f)H_1^*(f)H_2^{-1}(f)H_3^{-1}(f) & \text{when uplink noise is dominant} \\ P^*(f)H_1^*(f)H_2^*(f)H_3^{-1}(f) & \text{when downlink noise is dominant} \end{cases}$$

In practice, though, the filters $H_1(f)$, $H_2(f)$, and $H_3(f)$ are ignored with the justification that they alter the signal only a little and, for $H_2(f)$, anyway, that it varies with time (Chapter 12) so does not have a unique transfer function to use, anyway. When these three filters are ignored, the detection filter is simply the matched filter for the pulse filter:

$$Q(f) = P^*(f), \text{ in most applications}$$

This is not perfect but very good. Sometimes, though, perhaps to keep consumer terminals inexpensive or to keep using legacy terminals, the detection filter is less optimal:

$$Q(f) = \text{a filter with about same bandwidth as } P^*(f), \text{ in some applications}$$

It is easy to see roughly why the filter matched to the pulse, namely $P^*(f)$, makes sense as the detection filter when the filters $H_1(f)$, $H_2(f)$, and $H_3(f)$ can be neglected.

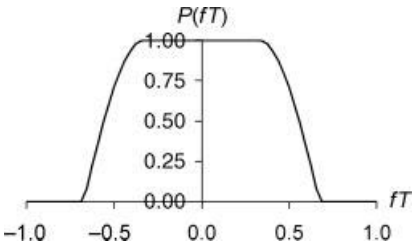


FIGURE 10.43 Transfer function of RRC pulse with $\alpha = 0.35$.

Let us look at the equivalent-baseband transfer function of the signal pulse, for example, that shown in Figure 10.43, which is the RRC pulse with $\alpha = 0.35$. Now, the purpose of the detection filter is to maximize the SNR prior to the symbol decision, so the detection filter must keep as much of the signal power and cut out as much of the noise power as possible. Where the pulse transform is flat, the detection filter should be flat and at the highest level, where the transform is zero the detection filter should be zero, and where it is in between the detection filter should be in between.

10.5.3.2 What Detection Filter Output from Signal Looks Like We want to see what the signal part of the detection-filter output looks like. The examples given are only for matched filters.

Since the pulse is the basic element of which the waveform for a symbol stream is constructed, we first look at the MF response to the pulse, namely $r(t)$ defined as $(p \circ q)(t)$ where $q(t) = p(-t)$. This is illustrated in Figure 10.44 for four pulses, the

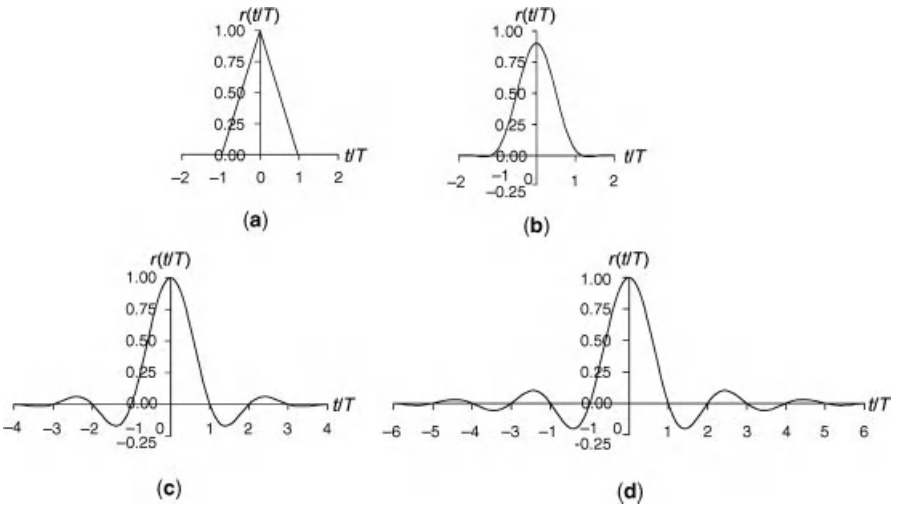


FIGURE 10.44 Matched filter output for pulse input: (a) rectangular pulse; (b) rectangular pulse with all but spectrum mainlobe filtered out; (c) RRC pulse with $\alpha = 0.35$; and (d) RRC pulse with $\alpha = 0.2$.

first two for the rectangular pulse and the last two for RRC pulses. The figure's second plot is actually for the variation on the rectangular pulse when only the mainlobe of the spectrum is kept.

The signal-only part $w(t)$ of the detection filter output is, for QPSK, BPSK, 8PSK, and 16QAM, in the equivalent-baseband representation, $\sum_k d_k r(t-kT)$, where d_k is the kth complex phasor (Section 10.2.1.2). For OQPSK and MSK, where I and Q bits are used sequentially, $w(t)$ is $\sum_k d_{kI} r(t-kT) + j d_{kQ} r(t-(k+\frac{1}{2})T)$, where d_{kI} and d_{kQ} are the kth I bit and Q bit, respectively. When $r(t)$ is zero at all nonzero multiples of T, then the value of $w(t)$ at any particular multiple of T has only one signal phasor contributing to it:

$$w(iT) = \sum_k d_k r(it-kT) = d_i r(0)$$

for the first group of modulations. There is one clear time at which $w(t)$ gives us information about a particular phasor. Similarly for the second group. This property, which the rectangular pulse and the RRC pulses have, means that there is no ISI out of the detection filter, where it matters the most (they do have ISI before the detection filter, where it matters less, as shown in Figure 10.31). The bandlimited rectangular pulse does not have this desirable property, that is, the signal has ISI.

A pulse has no ISI out of its matched filter if and only if the pulse satisfies **Nyquist's (first) criterion** (see, e.g., Proakis and Salehi (2008) and Benedetto et al. (1987)). The latter reference has a vivid drawing of what fulfilling the criterion means. $|P(f)|^2$ is chopped into segments R_s wide and they are all overlaid and added up, and the result is flat on a width R_s:

$$\sum_k \left|P\left(f-\left(k+\tfrac{1}{2}\right)R_s\right)\right|^2 = \text{brickwall filter on } -R_s/2 \text{ to } +R_s/2$$

We look at segments of matched-filtered signal corresponding to different symbol streams, to see generally what the signal looks like and to notice the no-ISI property of appropriate pulses. Figure 10.45 shows a BPSK signal with rectangular pulse, for three different infinite-length bit sequences. The independent axis is time. The waveform is sampled once per T, that is, at rate R_s. In each plot in the figure, the dashed lines are at the optimal **sampling times**. These times are optimal because the absolute value of the signal level is maximum there, averaged long-term. Indeed, the absolute value of the samples is the same in all three plots, which is due to the lack of ISI.

Figure 10.46 shows a similar result for the RRC pulse with $\alpha = 0.35$. Once again, the absolute value of a sample at the optimal sampling time does not depend on the surrounding bits.

The **eye pattern** is a plot of the I or Q component of the detection-filter output in time, like the plots above but with many consecutive symbols' worth of the filter output overlaid. The optimal sampling time can be seen; it is the time at which the eye is open the widest. There the signal part of the detection-filter output has the greatest amplitude, so the SNR is greatest. Checking for an open eye is a quick way

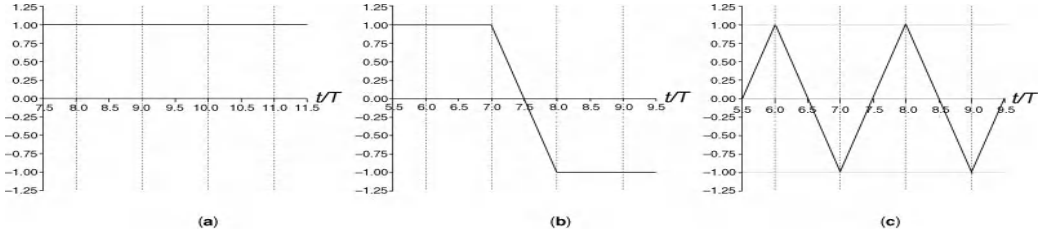


FIGURE 10.45 Segment of matched filter output of BPSK signal with rectangular pulse, for various bit sequences. (a) Bit sequence is $\ldots +1, +1, +1, +1 \ldots$; (b) bit sequence is $\ldots +1, +1, +1, -1, -1, -1 \ldots$; and (c) bit sequence is $\ldots +1, -1, +1, -1 \ldots$. Optimal sampling times are marked by vertical dashed lines.

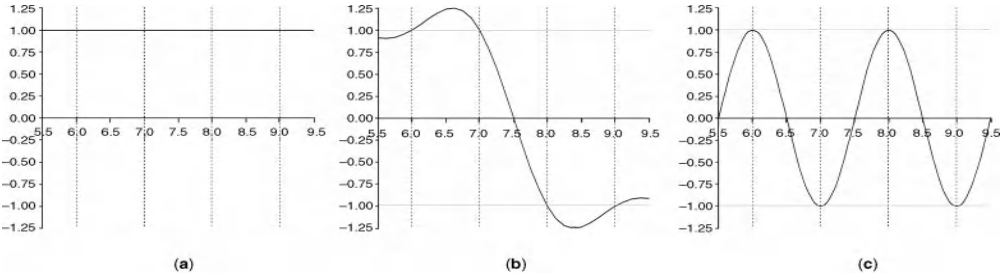


FIGURE 10.46 Segment of matched filter output of BPSK signal with RRC $\alpha = 0.35$ pulse, for various bit sequences. (a) Bit sequence is $\ldots +1, +1, +1, +1 \ldots$; (b) bit sequence is $\ldots +1, +1, +1, -1, -1, -1 \ldots$; and (c) bit sequence is $\ldots +1, -1, +1, -1 \ldots$.

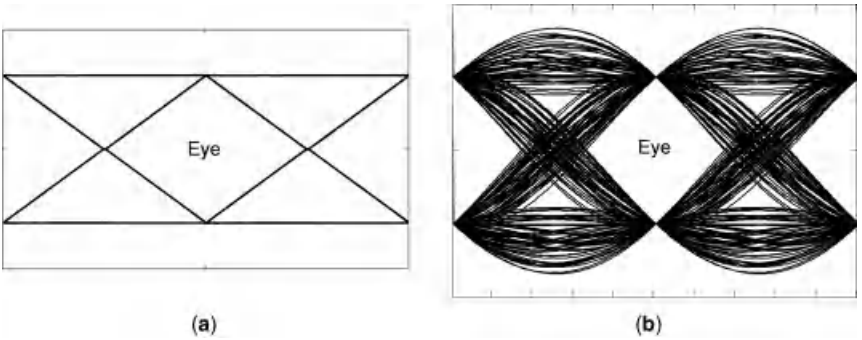


FIGURE 10.47 Example of BPSK eye patterns without noise for (a) rectangular pulse and (b) RRC $\alpha = 0.35$ pulse.

to verify that the transmitter and the receiver through the detection filter are probably working properly. The pattern is taken with a digital signal analyzer. Figure 10.47 gives examples of such patterns for BPSK without noise, for the rectangular pulse and the RRC $\alpha = 0.35$ pulse. Each pattern shows one eye in the middle surrounded by two half eyes. Figure 10.48 is similar but is for E_b/N_0 of 9.6 dB.

The detection filter removes spurious signals that are near-out-of-band or in nearby channels. It does not remove inband spurious signals.

10.5.3.3 Timing Recovery The sampling time in a symbol duration, relative to some defined beginning of symbol, is the **epoch**. **Timing recovery** or symbol synchronization tracks the epoch and perhaps the symbol rate R_s and provides them to the sampler. It is not necessary to track the derivative of symbol rate since this would have been removed by the carrier-tracking loop. (Doppler affects both carrier frequency and symbol rate by the same proportion.) Even with the carrier frequency tracked, there is still a small symbol rate variation left to track because in the transmitter the clock and oscillator are derived from different sources.

Timing recovery is conceptually similar to carrier recovery. In timing recovery, R_s corresponds to carrier frequency and timing jitter corresponds to phase noise.

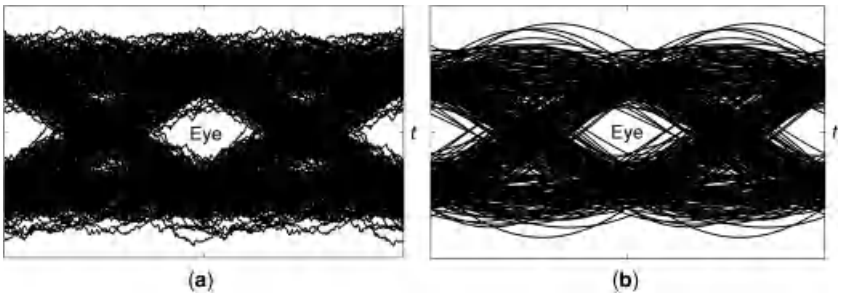


FIGURE 10.48 Example of BPSK eye patterns for E_b/N_0 of 9.6 dB and (a) rectangular pulse and (b) RRC $\alpha = 0.35$ pulse.

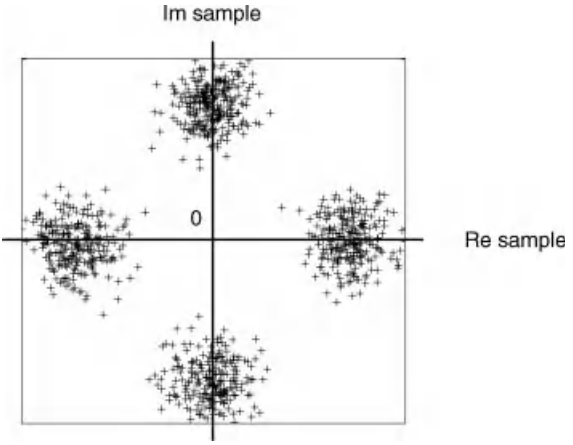


FIGURE 10.49 Example of QPSK scatter plot for E_b/N_0 of 9.6 dB and RRC pulse with $\alpha = 0.35$.

The only sources of timing jitter are the transmitter clock and the timing recovery loop's voltage-controlled clock, when the payload is bent-pipe. For more information see Proakis and Salehi (2008).

10.5.3.4 Sampling The input to the sampler is the detection-filtered, corrupted signal. Out of the sampler, for QPSK, BPSK, 8PSK, 16QAM, and 16APSK, the equivalent-baseband stream of corrupted phasors is $\sum_k \tilde{d}_k \delta(t - kT)$. A **phasor diagram** of sampled QPSK phasors in noise is shown in Figure 10.49. For OQPSK and MSK, where I and Q bits are used sequentially, the sampler output is $\sum_k \tilde{d}_{kI} \delta(t - kT) + j\tilde{d}_{kQ} \delta(t - (k + \frac{1}{2})T)$.

10.5.4 Symbol Decision Element

In some systems, symbol decision is performed and in others not. Most systems employ coding but directly input the samples of the detected noisy symbols into the decoder; such a decoder is "soft-decision." A "hard-decision" decoder does employ symbol decision, and so of course does an uncoded system. All are shown in Figure 10.50.

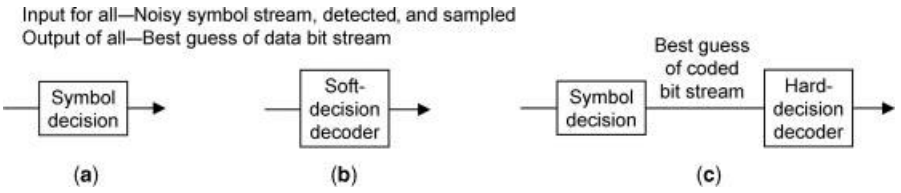


FIGURE 10.50 Data bit stream recovery: (a) uncoded data, (b) coded data with soft-decision decoding (most common), and (c) coded data with hard-decision decoding.

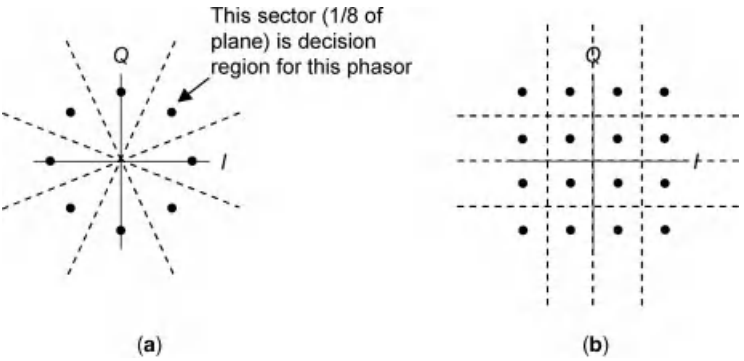


FIGURE 10.51 Decision regions for (a) 8PSK and (b) 16QAM.

The input to symbol decision is the stream of received, corrupted phasors and the output is the best guess of the (not yet decoded) transmitted bit stream. The output bit stream is actually a sequence of groups of bits where each group corresponds to a detected phasor. So the operation is first to decide what the transmitted phasor probably was and second to undo the assignment of bit groups to phasors.

The method of modulation-symbol decision is similar for all types of modulation we are treating. Let us start off by thinking about it for 8PSK and 16QAM. The complex plane is to be imagined, populated by the ideal phasors of the modulation scheme. The plane is divided into equal sectors that all start from the origin and contain a phasor. The edge between two sectors is equidistant from the two phasors in the sectors. The sectors are the **decision regions**. Now, whatever sector the received phasor falls into, the decision device decides that the ideal phasor in that sector was the one that was transmitted. These regions are illustrated in Figure 10.51 for 8PSK and 16QAM. For QPSK, BPSK, and 16APSK, it is the same idea. For OQPSK and MSK, it is a little different because the I and Q bits do not have transition at the same time, so their I and Q components can be thought of separately.

In digital communications systems, the quality measure of this recovered symbol stream is the **symbol error rate (SER)**. The quality measure of the bit stream directly recovered from the recovered symbol stream is the bit error rate (BER), specifically the **channel BER** or **uncoded BER**. The term "channel BER" uses "channel" in the communications theory sense, and the term "uncoded BER" refers to the fact that decoding and the resultant BER improvement have not occurred yet.

When the SER is low, it is almost always the case that when a symbol is detected wrong, the transmitted symbol was a nearest neighbor of the detected symbol (note that for 8PSK each symbol has two nearest neighbors, while for 16QAM some symbols have four, some have three, and the corner phasors have two). When the bits-to-symbol mapping was done with Gray coding, there is only one bit in error (Section 10.3.3.2):

$$\text{Uncoded BER} \approx \frac{R_s}{R_b}\text{SER} \quad \text{for low SER}$$

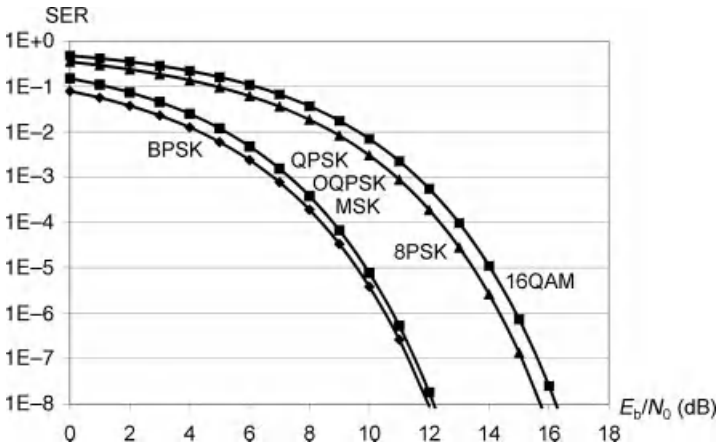


FIGURE 10.52 SER versus E_b/N_0 for AWGN channel.

Plots of SER versus E_b/N_0 are given in Figure 10.52 for all modulation schemes treated in the book, where the only signal corruption is AWGN.

10.5.5 Decoder

The decoder makes a best guess of what the original data bits were, based on the recovered "channel bit" stream. See, for example, Proakis and Salehi (2008) for full descriptions of decoders. If the uncoded BER is good enough, the **coded BER** is much better than the uncoded BER. Usually there is a direct relationship between the coded BER and the uncoded BER, that is, theoretical curves can be used. Because the encoder added redundancy to the data bits, the bit rate out of the decoder is lower than the channel bit rate R_b and equal to the original data bit rate.

Convolutional codes are decoded with the elegant **Viterbi algorithm**. First of all, if the code was punctured then the missing bits are replaced by zeros (while present bits are either -1 or $+1$). Let us suppose that the convolutional encoder divided the input data bit stream into k parallel bit streams and output i bits for each k in. The input bit stream can then be thought of as a sequence of groups of k bits, and the coded bit stream can be thought of as a sequence of groups of i bits. While all input sequences are possible, not all output sequences are possible. The Viterbi algorithm constructs all possible output sequences, called "paths," and it decides that the transmitted output sequence was the possible one that is closest to the received one. It outputs the sequence of input bit groups that corresponds to the output sequence it has decided on. By truncating how far back it looks for the most likely path, the Viterbi algorithm makes its output with a small fixed delay. The Viterbi algorithm can be implemented with either hard-decision or soft-decision input (no symbol decision).

The block decoders are similar in the sense that they also decide that the block sent was the one closest to the block received. The decoder outputs the block of bits

that would have been input to the encoder to produce the decided-on transmitted block. There are also both hard-decision and soft-decision implementations of decoders (Proakis and Salehi, 2008).

When a concatenated code is used, the decoding of the two codes is sequential, with deinterleaving in between.

Turbo decoding is different from the others in that decoding is iterative with the number of necessary iterations not known beforehand. Turbo decoding has soft-decision input. Typically, four iterations is enough if the achieved BER is 10^{-6}–10^{-5}, while 8–10 may be needed at BER of 10^{-5} (Proakis and Salehi, 2008).

10.6 SNR, E_S/N_0, AND E_B/N_0

The related terms SNR, E_s/N_0, and E_b/N_0 can cause confusion. We give their general definition and then specifically their values out of the sampler.

10.6.1 General Definition

The **signal-to-noise ratio (SNR)** is the ratio C/N of the signal power C to the noise power N at the output of a given filter. Sometimes "P" or "S" is used instead of "C." The filter can be actual or virtual, for example, we may want to know what the ratio would be if a particular filter were placed at a particular point.

SNR can be measured with a power meter at filter output. Alternatively, it can be calculated as follows. Suppose the signal is $x(t)$ as a function of time, the impulse response of the filter of interest is $h(t)$, $y(t)$ is the virtually filtered signal, and the noise power spectral density is N_0 over the filter's passband. Then the calculated signal power and noise power are, respectively,

$$C = \int \mathcal{S}_y(f)df = \int \mathcal{S}_x(f)|H(f)|^2 df \text{ and } N = N_0 \int |H(f)|^2 df$$

If the noise psd is not flat across the filter's passband then the more general equation for noise power must be used:

$$N = \int \mathcal{S}_n(f)|H(f)|^2 df \quad \text{where } \mathcal{S}_n(f) = \text{noise psd as a function of frequency}$$

Another way to compute SNR, a rougher way, is often used in signal-and-noise level budgets (Section 8.3). The estimated unfiltered signal power at a point in the signal path, the estimated N_0, and the noise bandwidth of an interesting filter are used.

$\boldsymbol{E_s/N_0}$ and $\boldsymbol{E_b/N_0}$ are respectively **symbol-energy-to-noise-psd** and **bit-energy-to-noise-psd**. They are more subtle terms than SNR because apparently no filter is involved in the definitions. However, no noise really has a psd that is constant over "all" frequencies, so N_0 must be taken over some frequency band, but which? And if the psd is not constant over this band then exactly how should it be averaged over

the band? To answer these questions we must call to mind how E_s/N_0 is used, and that is primarily as the independent variable in SER or BER plots. Now, in the plots of theoretical SER or BER over an AWGN channel, E_s/N_0 is the SNR out of the matched filter at the optimal sampling time. It also equals $(C/R_s)/N_0$. However, on a channel with some other filters and a detection filter that is not the matched filter, these two quantities are not equal, as we will see in the next subsection. Also, as we have noted, the noise psd may not be constant. The truest definition of E_s/N_0 is the SNR out of the detection filter at the sampling time, which maximizes the SNR, if the detection filter were placed at the point of interest. A more easily computable definition of E_s is C/R_s where C is unfiltered signal power at the point of interest, and N_0 is a weighting of the noise psd by the detection filter $P(f)$:

$$N_0 = \int \mathcal{S}_n(f)|P(f)|^2 df \Big/ \int |P(f)|^2 df$$

Once E_s/N_0 is computed, then E_b/N_0 can be computed as $(E_s/N_0)(R_s/R_b)$.

10.6.2 Values at Sampler Output

We want to compare different cases of E_s/N_0 at the sampler output. Let us start with the ideal case, where there is no signal distortion besides AWGN and the detection filter is the matched filter of a pulse which satisfies the Nyquist criterion. For QPSK, BPSK, 8PSK, 16QAM, and 16APSK, the equivalent baseband signal $x(t)$ into the matched filter is

$$x(t) = \sum_k d_k p(t - kT) + n(t)$$

We now filter this by the matched filter to get $y(t)$:

$$y(t) = \sum_k d_k \int p(\tau - kT)p(\tau + t)d\tau + \int n(\tau)p(\tau + t)d\tau$$

Sampled at the optimal sampling time zero for d_0 this is

$$y(0) = d_0 \int (p(\tau))^2 d\tau + \int n(\tau)p(\tau)d\tau$$

The following is proportional to its power:

$$E|y(0)|^2 = E\left(|d_0|^2\right)\left[\int (p(\tau))^2 d\tau\right]^2 + \int d\tau \int dt\, E[n(\tau)n^*(t)]p(\tau)p(t)$$

We note that $E|d_0|^2 = E|d_k|^2$ for any k.

In order to evaluate the second term on the right-hand side, we have to introduce the **autocorrelation function** $\boldsymbol{\mathcal{R}(t)}$ of a random process. To show that $\mathcal{R}(t)$ applies to a specific random process $z(t)$, it would be written as $\mathcal{R}_z(t)$. When $z(t)$ fulfills

some conditions, then $\mathcal{R}(\tau)$ is defined as $E[z(\tau+t)z^*(t)]$ where E is the expected value (Papoulis, 1984). Usefully, the spectrum $\mathcal{S}(f)$ of $z(t)$ is just the transform of $\mathcal{R}(t)$: $\mathcal{S}(f) = [\mathcal{F}(\mathcal{R})](f)$ where $\mathcal{F}$ is Fourier transform.

Back to our computation. The autocorrelation function of white Gaussian noise is the scaled delta function $N_0\delta(t)$. So

$$E|y(0)|^2 = E\left(|d_0|^2\right)\left(\int (p(t))^2 dt\right)^2 + \int N_0 (p(t))^2 dt$$
$$= E|d_0|^2 T^2 + N_0 T \quad \text{where} \quad p(t) \text{ has been scaled so that } \int (p(t))^2 dt = T$$

The E_s/N_0 out of the sampler is then $E|d_0|^2/R_s/N_0$.

We now look at a nonideal case, where the detection filter is matched to the same pulse as above but there has been non-negligible filtering $H_1(f)$, $H_2(f)$, and $H_3(f)$ (see Figure 10.13), so the filtered pulse is $\tilde{p}(t) \triangleq [p \circ h_1 \circ h_2 \circ h_3](t)$. Now the signal out of the detection filter $P(f)$ is

$$y(t) = \sum_k d_k \int \tilde{p}(\tau - kT) p(\tau + t) d\tau + \int n(\tau) p(\tau + t) d\tau$$

and, if 0 is still the optimal sampling time for d_0,

$$E|y(0)|^2 = E\left(|d_0|^2\right)\left[\sum_k \int \tilde{p}(\tau - kT) p(\tau) d\tau\right]^2 + \int d\tau \int dt\, E[n(\tau) n^*(t)] p(\tau) p(t)$$
$$= E\left(|d_0|^2\right)\left[\sum_k \int \tilde{p}(\tau - kT) p(\tau) d\tau\right]^2 + N_0 T$$

The E_s/N_0 out of the sampler in this case is then

$$E_s/N_0 = E\left(|d_0|^2\right)\left|R_s \sum_k \int \tilde{p}(\tau - kT) p(\tau) d\tau\right|^2 / R_s/N_0$$

For comparison, E_s/N_0 computed right before the detection filter is something different:

$$E_s/N_0 = E\left(|d_0|^2\right)\left(R_s \int (\tilde{p}(t))^2 dt\right)^2 / R_s/N_0$$

10.7 SUMMARY OF SIGNAL DISTORTION SOURCES

Table 10.1 summarizes the sources in the end-to-end communications system of the various types of signal distortion besides additive noise. The information is for a bent-pipe payload but it is obviously alterable for a processing payload.

TABLE 10.1 Summary of Signal Distortions in Bent-Pipe Payload Satellite Communications System

Signal Distortion Type	Occurs in Ground Transmitter	Occurs in Payload	Occurs in Ground Receiver	Source	Mechanism	Reference Sections
Clock jitter	Yes	No	Yes	Clock	Symbol timing and timing recovery	10.5.3.3
Linear distortion	Maybe	Yes	Maybe	Filters	Input and output multiplexing; legacy detection filter not matched to pulse	7.2 and 7.3; 10.5.3.1
Phase noise and spurious phase modulation	Yes	Yes	Yes	Oscillators	Modulation; frequency conversion; demodulation	7.5; 10.5.2
Long-term frequency instability	Perhaps in consumer equipment	Yes	Perhaps in consumer equipment	Oscillators	Aging	7.6
Frequency converter-caused spurious signals	Yes	Yes	Yes	Frequency converters	LO harmonics, signal harmonics, cross-products of signal and LO, leakage of input signal and LO	7.7
Inband and near-inband intermodulation products	Maybe	Yes	No	HPA	Bandpass nonlinearity	7.8
HPA-caused spurious signals	Maybe	Yes	No	HPA and its power supply	Various	7.9

APPENDIX 10.A

10.A.1 Sketch of Proof That Pulse Transform and Signal Spectrum Are Related

We invoke probability theory to sketch the proof of an important property of the spectrum of the complex baseband signal $z(t)$, which we consider to be a random or stochastic process (Papoulis, 1984).

For QPSK, 8PSK, 16QAM, and 16APSK, the baseband signal $z(t)$ can be written as $\sum_k d_k p(t-kT)$ where d_k is the kth complex phasor and $p(t)$ is the real-valued pulse function. We assume that the phasors d_k of the data sequence are chosen independently of each other and that each is equally likely to be any one from the phasor collection. Then $\mathcal{R}(\tau)$ is as follows:

$$\begin{aligned}\mathcal{R}(\tau) &= E\left[\sum_k d_k p(\tau+t-kT)\sum_i d_i^* p(t-iT)\right]\\ &= \sum_k E\left[d_k d_i^*\right] p(\tau+t-kT)\sum_i p(t-iT)\\ &= \sum_k E\left[|d_k|^2\right] p(\tau+t-kT)p(t-kT)\\ &= E\left[|d_k|^2\right][p(t)\circ p(-t)](\tau)\end{aligned}$$

The only terms that matter in the product of the two sums above are those for which $k=i$; when $k\neq i$, $E(d_k d_i^*)=0$ because d_i is equally likely to be any phasor or its negative.

For OQPSK and MSK, the baseband signal $z(t)$ can be written as follows:

$$z(t) = \sum_k d_{kI} p(t-kT) + j d_{kQ} p(t-\left(k+\tfrac{1}{2}\right)T)$$

where d_{kI} and d_{kQ} are the kth I bit and Q bit, respectively. We assume that the bits are all independent of each other and that each is equally likely to be -1 or 1. Then $\mathcal{R}(\tau)$ for $z(t)$ is as follows:

$$\begin{aligned}\mathcal{R}(\tau) &= E\left\{\sum_k \left[d_{kI}p(t-kT)+jd_{kQ}p(t-(k+\tfrac{1}{2})T)\right]\sum_i \left[d_{iI}p(t-iT)+jd_{iQ}p(t-(i+\tfrac{1}{2})T)\right]\right\}\\ &= \sum_k \left(Ed_{kI}^2\right)p(\tau+t-kT)p(t-kT)\\ &\quad +\sum_k \left(Ed_{kQ}^2\right)p(\tau+t-(k+\tfrac{1}{2})T)+p(t-(k+\tfrac{1}{2})T)\\ &= E\left[|d_k|^2\right][p(t)\circ p(-t)](\tau)\end{aligned}$$

where the d_k again are the possible complex phasors. So the result is the same as for the other modulation schemes.

We invoke properties of the Fourier transform, where $u(t)$ and $v(t)$ are any two time functions:

$$[\mathcal{F}(u \circ v)](f) = U(f)V(f) \quad \text{and} \quad [\mathcal{F}(v^*(-t))](f) = V^*(f)$$

and we find the simple relationship between the signal spectrum and pulse transform for all of these modulation schemes:

$$\mathcal{S}(f) = E\left[|d_k|^2\right]|P(f)|^2 \quad \text{where } E\left[|d_k|^2\right] \text{ is the average phasor power.}$$

REFERENCES

Benedetto S, Biglieri E, and Castellani V (1987). *Digital Transmission Theory*. Englewood Cliffs, New Jersey: Prentice-Hall, Inc.

Couch II LW (1990). *Digital and Analog Communication Systems*, 3rd ed. New York: Macmillan Publishing Company.

Elbert BR (2004). *The Satellite Communication Applications Handbook*, 2nd ed. Boston: Artech House, Inc.

ETSI EN 302 307, v1.2.1 (2009). Digital video broadcasting (DVB); second generation framing structure, channel coding and modulation systems for broadcasting, interactive services, news gathering and other broadband satellite applications (DVB-S2). Sophia-Antipolis Cedex, France: European Telecommunications Standards Institute.

Gardner FM (1979). *Phaselock Techniques*, 2nd ed. New York: John Wiley & Sons.

Lin S and Costello Jr. DJ (1983). *Error Control Coding, Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Maral G and Bousquet M (2002). *Satellite Communications Systems*, 4th ed. Chichester, England: John Wiley & Sons Ltd.

Papoulis A (1984). *Probability, Random Variables, and Stochastic Processes*, 2nd ed., International Student ed. Singapore: McGraw-Hill, Inc.

Proakis JG and Salehi M (2008). *Digital Communications*, 5th ed., International ed. New York: McGraw-Hill.

CHAPTER 11

COMMUNICATIONS LINK

11.1 INTRODUCTION

This chapter describes the remaining elements of the end-to-end satellite communications system not discussed in earlier chapters, namely the communication links. The payload has been extensively described in Part I of this book. The modulating transmitter and the demodulating receiver, minus the antennas, have been described in Chapter 10. In this chapter, we look at what happens to the signal when it goes between a ground transmitter and the satellite (the **uplink**) and between the satellite and a ground receiver (the **downlink**). Some of the effects happen on the link, and some happen only as the transmitted signal is received by the antenna. The payload is assumed to be bent-pipe but it should be rather easy to see how things would be for a processing payload.

The rest of this chapter is organized as follows:

- *Section 11.2*: End-to-end C/N_0
- *Section 11.3*: Signal power on link: what determines it and what makes it vary, including atmospherics
- *Section 11.4*: Noise level on link: what determines it, including antenna noise, and what makes it vary
- *Section 11.5*: Interference on link: how it can arise and what mitigates it
- *Section 11.6*: Link budget, combining these facets.

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

11.2 END-TO-END C/N_0

At this point neglecting interference, the quality of the (impaired) signal on the end-to-end communications system is represented to first order by the C/N_0 out of the ground receiver's RF/IF front end. C is signal power (sometimes denoted by "P" or "S") and N_0 is the RF one-sided power spectral density (psd) of the system noise (Section 7.12). Whether this quality number is good enough is told, approximately, by comparison with the C/N_0 needed by the receiver's baseband part.

We show that by treating the payload and the ground receiver's front end as electronic elements with gains and noise figures and following the C/N_0 calculation through uplink and downlink, we obtain a composite C/N_0 that conforms to the well-known formula. We assume that the noise on the links is additive white Gaussian noise (AWGN). Any noise added to the signal in the ground transmitter is negligible compared to payload receiver noise.

Figure 11.1 shows the computation of uplink C/N_0 from a noise temperature T_{a1} and the payload's gain G_1 and noise figure F_1. T_{a1} is the payload's antenna-noise temperature T_a, as defined in Section 7.12. Also, T_{a1} plus the payload's electronics-noise temperature equals the payload's system noise temperature.

Figure 11.2 shows the computation of downlink C/N_0 from a noise temperature T_{a2} and the station's gain G_2 and noise figure F_2. T_{a2} is the ground station's

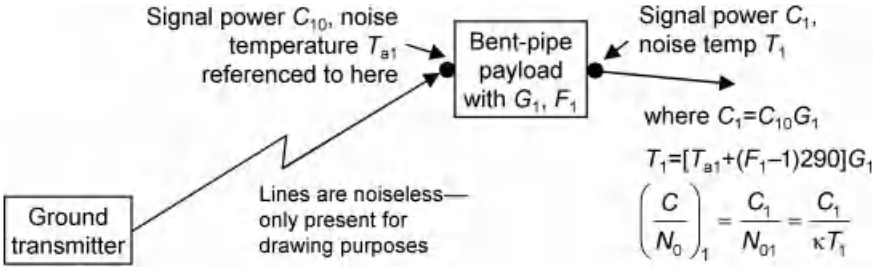


FIGURE 11.1 Uplink C/N_0 derived from payload gain and noise figure.

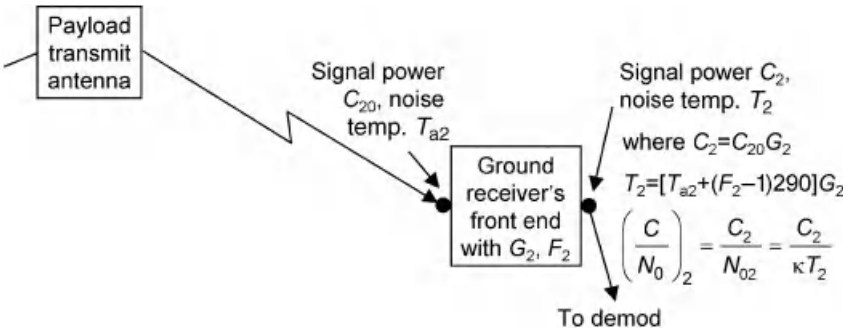


FIGURE 11.2 Downlink C/N_0 derived from ground receiver gain and noise figure.

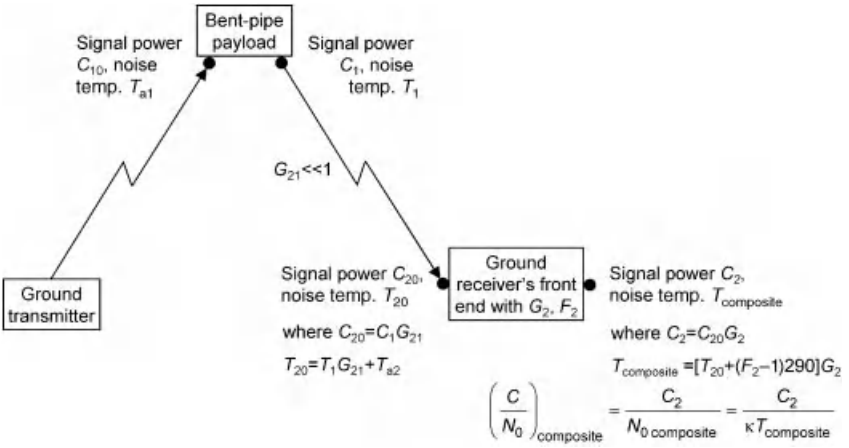


FIGURE 11.3 Composite C/N_0 derived by combining link gains and noise figures.

antenna-noise temperature T_a. Also, T_{a1} plus the ground station's electronics-noise temperature equals the ground station's system noise temperature.

The important number is the composite C/N_0, after the signal has gone through both links. C is the power of the signal component of what is received, and N_0 is the power spectral density due to all noise sources on both links. Figure 11.3 shows the computation of composite C/N_0 from the antenna noise temperatures and the composition of gains and noise figures. Downlink propagation is equivalent to a big loss without noise figure. We need to show that the composite C/N_0 represented in the figure conforms to the well-known formula for the composition of the uplink and downlink C/N_0's

$$\left(\frac{C}{N_0}\right)^{-1}_{\text{composite}} = \left(\frac{C}{N_0}\right)^{-1}_{1} + \left(\frac{C}{N_0}\right)^{-1}_{2}$$

From the figure, we have

$$\left(\frac{T}{C}\right)_{\text{composite}} = \frac{T_{\text{composite}}}{C_2} = \frac{T_1G_{21}G_2 + T_2}{C_2} = \frac{T_1}{C_2/(G_{21}G_2)} + \frac{T_2}{C_2} = \frac{T_1}{C_1} + \frac{T_2}{C_2}$$

So it is correct.

11.3 SIGNAL POWER ON LINK

11.3.1 Introduction

We wish to look at what determines the signal power on the link. There are factors that are constant and factors that vary. Overviews of the two sets of factors are given in Tables 11.1 and 11.2, respectively, for the downlink (the uplink would be

TABLE 11.1 Constant Losses in Downlink

Loss Type	Source	Sections
Average loss from payload antenna pointing error	Spacecraft pointing error (worse during station-keeping maneuvers); misalignment of antenna as installed on spacecraft	11.3.2
Average free-space loss	Inversely proportional to distance squared	11.3.3
Attenuation by atmospheric gas	Oxygen and water vapor, for carrier frequency above 10 GHz	11.3.6
Loss from radome, reflector, and feed cover at BOL	–	N/A
Average loss from ground antenna pointing error	–	N/A
Average polarization-mismatch loss	Polarization ellipses different and/or in different orientations, of payload and ground antennas	11.3.7

TABLE 11.2 Varying Losses in Downlink

Variation Type	Source	Sections
Payload power-out instability (preamp in ALC mode)	Drift in preamp output power over life, variation of HPA's P_{out} with temperature, drift in HPA's P_{out} over life	7.10
Payload gain instability (preamp in fixed-gain mode)	Variation of HPA's P_{out} with temperature, drift in HPA's P_{out} over life	7.10
Jump in transmitted power	Switch over to redundant HPA (more switches and waveguide or coax post-HPA)	6.4.2.3 and 6.5.1.2
Payload antenna pointing error	Spacecraft pointing error (worse during station-keeping maneuvers)	11.3.2
Varying free-space loss	Non-GEO (chiefly) varying distance to ground station	11.3.3
Varying atmospheric attenuation	Ionospheric or tropospheric effects, depending on carrier frequency	11.3.4–11.36
Varying ground antenna loss	Water or snow on feed, reflector, or radome; aging	Crane, 2002; Crane and Dissanayake, 1997
Varying ground antenna mispointing	–	N/A
Varying polarization-mismatch loss	–	N/A
Jump in front-end loss	Switch over to redundant LNA or back to primary (different number of switches and more or less waveguide or coax)	N/A

similar). The last column in the tables gives references for the factors except for those belonging to the ground station. Each end of a link must keep the variations it causes within bounds and contribute to accommodating the total variations. In the rest of this section, we discuss the factors in payload-to-ground order.

11.3.2 Loss and Variation from Payload Antenna Pointing Error

The payload antenna has a slight fixed component of mispointing, due to misalignment upon installation on the spacecraft. To any particular place on the earth this will cause either a small increase or decrease in gain relative to the no-misalignment gain.

However, larger than this is the variation in antenna gain due to antenna pointing error, which in turn is due to spacecraft pointing error. Antenna pointing error has two physically orthogonal but statistically dependent dimensions, with generally different nonzero means and different standard deviations (Maral and Bousquet, 2002). In addition, the pointing error has two states, during thrusting and otherwise; during thrusting the standard deviations of the error's two dimensions are larger than otherwise. Antenna gain is a function of the two-dimensional pointing error. The gain is not necessarily symmetric about zero pointing error and in most cases has a nonzero average loss. Nor does the gain necessarily have the same characteristics in the two dimensions.

11.3.3 Free-Space Loss

Free-space loss is basically the fraction of EIRP that actually arrives at the receiving antenna. The EIRP spreads out from a point in the transmit antenna (the phase center—Section 3.2.3.2) at the speed of light to the huge spherical "surface" that contains the receive antenna. The surface area of this sphere is $4\pi R^2$ where R is the distance between the payload and the ground station. What matters, actually, is the surface area in terms of the carrier wavelength, that is, $4\pi(Rf/c)^2$ where f is carrier frequency and c is the speed of light in vacuum. With some convenient scaling, the free space "gain" is defined as $[c/(4\pi Rf)]^2$. In dB, it would be 10 times the log of this.

It is evident that the free-space loss in dB is proportional to frequency in dBHz. For a very wide channel (on the order of 15% of center frequency), the resultant gain slope across the band will be significant. Fortunately, this gain slope is canceled by the gain slope of the transmit (or receive) antenna.

11.3.4 Loss and Variation from Atmospheric Attenuation

The effects of the atmosphere on the signal are different depending on the signal's frequency band. For frequencies between 1 and 10 GHz, ionospheric scintillation and the Faraday effect are the only important atmospheric impairments. For frequencies above 10 GHz, ionospheric effects are negligible and the dominant atmospheric effect is rain attenuation, followed by atmospheric gas attenuation (Recommendation, ITU-R, P.618-10).

11.3.5 Atmospheric Effects for Carrier Frequencies from 1 to 10 GHz

The only significant atmospheric impairments for carrier frequencies below 10 GHz come from the ionosphere (Recommendation, ITU-R, P.618-10), the ionized layer of the atmosphere between about 15 and 400 km altitude (Ippolito, 2008). The important impairments are scintillation and the Faraday effect, which are caused by different aspects of the ionosphere (Recommendation, ITU-R, P.531-10).

The worse effect is scintillation, caused by fluctuations in the refractive index due to small-scale inhomogeneities of the ionosphere (Recommendation, ITU-R, P.531-10). It makes a signal's amplitude and phase fluctuate rapidly about the nominal value (Allnutt, 1989). It affects L-band communications, for example, those of Inmarsat, Iridium, and Globalstar, as well as S-band but C-band insignificantly. There are three zones in which it occurs, within 20° of the magnetic equator and near the two magnetic poles. A scintillation event can last from half an hour to several hours. For equatorial ground stations in years of solar maximum, scintillation occurs almost every evening (Recommendation, ITU-R, P.531-10). In other years, it is strong within a month of the equinoxes (Davies and Smith, 2002). The fading rate is about 0.1–1 Hz (Recommendation, ITU-R, P.531-10), that is, slow enough to be followed by an automatic gain-control (AGC) unit with enough range. Phase fluctuations could be tracked by a phase-locked loop. Table 11.3 shows amplitude and phase scintillation values for a bad solar condition but not the worst. These values, as well as the ones for Faraday rotation, come from the model given in Recommendation, ITU-R, P.531-10.

Faraday rotation was mentioned in Section 3.2.3.4 in regard to antenna polarization selection. Here we explain it. The rotation is directly proportional to the integral of the product of the ionosphere's electron density and the component of the earth's magnetic field along the signal path (Recommendation, ITU-R, P.531-10). A linearly polarized signal splits into two circularly polarized signals, which travel with slightly different speeds and along slightly different paths. When they recombine the linear polarization has been rotated (Davies and Smith, 2002). Circular polarization is not affected. The rotation is large at L-band, small but significant at C-band, and insignificant at 10 GHz as can be seen from Table 11.3. Faraday rotation has relatively regular diurnal, seasonal, and 11-year solar-cycle variations and strongly depends on geographical location (Recommendation, ITU-R, P.531-10).

TABLE 11.3 Estimated Ionospheric Effects Near Equator, Elevation About 30°

Carrier Frequency (GHz)	Scintillation[a] (dB pk-pk)	Scintillation[a] (° rms)	Faraday Rotation[b] (°)
1	>20	>45	108
3	≈10	≈26	12
10	≈4	≈12	1.1

Source: Recommendation, ITU-R, P.618-10.

[a]Values observed near geomagnetic equator during early night-time hours at equinox under conditions of high sunspot number.

[b]Based on a high (but not highest) TEC value encountered at low latitudes in daytime with high solar activity.

11.3.6 Atmospheric Effects for Carrier Frequencies Above 10 GHz

For frequencies between about 10 and 30 GHz, the only important atmospheric effects on the signal power are from rain and, to a smaller extent, gas. For both effects, whenever the attenuation along a zenith path is given, it can be converted to another elevation greater than 5% by multiplying it by the cosecant of the elevation angle (Recommendation, ITU-R, P.618-10; Recommendation, ITU-R, P.676-7, respectively).

11.3.6.1 Rain Attenuation Rain attenuation can be a complicated matter to treat in payload design when the coverage area is large because the range of expected attenuations over the coverage area can be large and dependent on several factors at each location.

Usually a satellite customer wants to have a very small long-term **link-outage** probability, around 0.01% for Ku-band and 0.1–1% for Ka-band (Feldhake, 1997). The link being "out" means that it does not "close." The **link closes** when the SNR into the receiver is sufficient for the link's required performance. Link closing is also called link **availability**. For a downlink, the payload may be called on to provide a **rain margin**, that is, extra power, to guarantee against most cases of rain at a ground station, that is, to provide long-term the desired link-outage probability. The customer specifies a particular probabilistic model of rain attenuation for the payload manufacturer to use. A rain model gives rain attenuation as a function of **exceedance** probability or vice versa, with the meaning that the rain attenuation will exceed a particular value for a given small percentage of the time. The rain margin is set to the attenuation value expected at the exceedance that equals the desired link-outage probability.

"Long-term" means over a period of years, as the rain varies from year to year. The ITU-R says that for a given exceedance in the range from 0.001 to 0.1%, the attenuation in dB varies from year to year with more than a 20% rms, for the same signal path, frequency, and polarization (Recommendation, ITU-R, P.618-10). The necessary period to average over to obtain reliable statistics depends on the characteristics of the local climate and usually is between 3 and 7 years (Recommendation, ITU-R, P.837-5).

A rain model has two parts, a rain-rate model and a rain-attenuation model, which in some cases can be separately selected. The rain-rate model provides the long-term rain-rate probability distributions (rate versus exceedance percentage or vice versa). For a given location, the distribution results either from long-term measurements at or near the location or is obtained from the model's collection of climate-zone maps and a list of zone properties. In the latter case, the distribution is adjusted for the location's altitude. The rain-attenuation model then factors in the carrier frequency, elevation angle to the satellite, and the polarization tilt angle with respect to the horizontal (Ippolito, 2008). The polarization tilt angle is the angle between the polarization ellipse's major axis and the local horizontal.

The two best-known rain models are the ITU-R Rain Attenuation Model and the Crane Global Model (Ippolito, 2008). The ITU-R model is an international

standard. Its rain-rate model was embodied in a set of worldwide climate zone maps and a list of zone properties until October 1999, when it changed to contour maps of the rain rate at the 0.01% exceedance level (Recommendation, ITU-R, P.837-5; Recommendation, ITU-R, P.618-10). The second part of the ITU-R model, namely the attenuation model, has since 1999 been based on the DAH rain-attenuation model, named for its authors (Dissanayake, Allnutt, and Haidara) (Ippolito, 2008). The ITU-R changed to this model after a study comparing 10 models (but not the Crane Global) against ITU-R propagation data showed that the DAH model provided the most consistent and accurate predictions (Feldhake, 1997). The model is applicable from 4 to 35 GHz and for exceedances from 0.001 to 10% (Dissanayake et al., 1997). It is easy to compute. The model takes as input the rain rate at the 0.01% exceedance level. The DAH model was developed from a fit to all the rain-attenuation data in the ITU-R propagation database, which at the time held (only) altogether 120 years of rain attenuation observations from various places around the world (Crane and Dissanayake, 1997).

The Crane Global Model is used frequently, especially in the United States (Feldhake, 1997). It was the first published model (1980) that provided a self-contained rain-attenuation prediction procedure for the entire world. The rain-rate model is embodied in a set of worldwide climate-zone maps and a list of zone properties. The attenuation model is based entirely on observations of rain rate, rain structure, and the vertical variation of atmospheric temperature, and not on attenuation measurements (Crane, 1996). It differs from the DAH model in being founded on the physics of rain attenuation (Crane and Dissanayake, 1997). Crane also has a Two-Component Model of rain rate and of rain attenuation, which he updated along with the Global Model in 1996 (Crane, 1996). The Two-Component Model uses two rain-rate distributions, one to describe the small (about 1 km) cells of intense rain and the much larger (roughly 30 km (Crane, 1993)) surrounding regions of lower intensity. All storms contain both types of rain but a given signal path may not intersect a high-intensity cell. These two types of rain areas correspond approximately to the convective rain cell and stratiform region, respectively, of other models (Crane, 2003a). Stratiform rain typically has stratified horizontal extents of hundreds of kilometers. Convective rain occurs in a cell with a horizontal extent of usually several kilometers and has convective upwelling (Ippolito, 2008). He also has a Local Model for rain rate, which is his initial Two-Component Model augmented by an adjustment for year-to-year variation (Crane, 2003a). At the time he published it, only in the United States was there enough rain-rate data to apply the model. The National Climate Data Center has been collecting measurement data at 105 US sites since 1965. The way to apply the model is to use the data for the nearest of those sites. The low-exceedance rain rates from that site are good enough to use because low-exceedance rain levels are caused by high-intensity rain cells, which are similar enough at both locations. Interestingly, local climate variations due to mountains, large water bodies, city heat islands, and so on are not important, for the same reason. The Crane models can also be found in Crane (2003b). Excel macros for both Global and Local are to be found on the Internet at weather.ou.edu/~actsrain/crane/model.html.

Crane compared models of rain rate against measurements for US-wide sites. The difference in rms error among the ITU-R, Crane Global, and Crane Local models were statistically insignificant, for exceedances from 0.001 to 1%, although relatively speaking the Local Model did the best and the Global Model did the worst. The average deviation of prediction from measurements was within 0.15 dB in all cases (Crane, 2003a).

The rain data in the ITU-R database does not uniformly represent all parts of the world. Most data has been taken by the United States and Europe. However, the 2007 version of the ITU-R rain-rate maps has incorporated long-term measurements of rain at 14 sites in Korea, China, and Brazil (Recommendation, ITU-R, P.837-5), while the previous (2003) version did not, so the ITU-R is continuing to add acceptable data to their database. There is comparatively little propagation data in the tropical region, which lies between the Tropic of Cancer and the Tropic of Capricorn, that is, within 23.4° of the equator (Green, 2004). However, the amount of data is increasing and models are being developed. A comparison of three years of measurements at 12 GHz in four tropical locations with five rain-attenuation models showed that the best model was that by Ramachandran and Kumar (2006), for exceedances from 0.001 to 1% (Mandeep and Allnutt, 2007). The ITU-R model predicted well except at exceedance about 0.001%, where it over-estimated. In a measurement campaign in Thailand, 5 years of rain rate and 3 years of attenuation measurements were taken at 13 GHz and compared with the ITU-R and Crane Global model predictions (Boonchuk et al., 2005). Even though the ITU-R model predicted the rain rate for exceedances from 0.001 to 1% well, interestingly it underestimated by a few dB the rain attenuation for exceedances from 0.01 to 1%. For the Crane Global Model, the site lies near the boundary of two climate zones. For one of these zones, the measured rain rate matched the prediction well. The measured rain attenuation lay between the predicted rain attenuations for the two zones. The Tropical Rainfall Measuring Mission satellite has been taking rainfall data in the tropics with radar since 2000. It is a joint mission of NASA and the Japan space agency JAXA (NASA, 2010). The rain in the tropics is different from the rain in the temperate zones; the rain in temperate zones is mostly stratiform, while in the tropics it is mostly convective. A signal path may intersect two rain cells instead of just one or none (Mandeep and Allnutt, 2007).

Let us use the ITU-R rain model to provide some information for rule-of-thumb calculations. First, we compute the "specific attenuation" for various frequencies and compare them. The specific attenuation is in units of dB/km and corresponds to an exceedance of 0.01%. To first order, the specific attenuation times the path length through rain equals the attenuation $A_{0.01}$ for the exceedance 0.01%. Table 11.4 gives the specific attenuation versus frequency for a typical temperate climate and for a typical tropical climate, for circular polarization. The numbers would not be very different for linear polarization (Ippolito, 2008). The specific attenuation at 30 GHz is roughly four times the value at 12 GHz, for both climates. So on the same path, under identical conditions except for frequency, the attenuation at 30 GHz would be about four times the attenuation at 12 GHz. The tropical climate's specific attenuation is roughly three times that of the temperate climate at each frequency. A second

TABLE 11.4 Specific Attenuation Versus Frequency for Typical Temperate and Tropical Climates, for Circular Polarization

Frequency (GHz)	Specific Attenuation in dB/km for Typical Temperate Climate ($R_{0.01} = 35$ mm/h)	Specific Attenuation in dB/km for Typical Tropical Climate ($R_{0.01} = 100$ mm/h)
12	1.5	4.9
15	2.2	6.9
20	3.5	10.3
30	6.4	17.1

set of interesting numbers are the rain attenuations $A_{0.1}$ and A_1 at 0.1 and 1% exceedances, respectively, scaled by $A_{0.01}$, for a given carrier frequency, latitude, and elevation at least 20° of a GEO satellite. The two ratios are plotted versus $A_{0.01}$ in Figure 11.4. The ratio $A_{0.1}/A_{0.01}$ has a sensitivity to ground-station latitude and elevation angle, which causes the spread of values shown in Figure 11.4, but the ratio $A_1/A_{0.01}$ does not. If you know $A_{0.01}$ you can compute A_1 and get a rough idea of $A_{0.1}$.

11.3.6.2 Other Atmospheric Effects and Combined Effects The other atmospheric effects we discuss are the attenuation due to atmospheric gases, clouds, and tropospheric scintillation.

In most cases, it is sufficient to model the attenuation due to gas as a constant loss, for each location of interest (Recommendation, ITU-R, P.618-10). The attenuation from absorption by oxygen is relatively constant at each location, while the attenuation from water vapor varies. The annual variation in the water vapor density is a Gaussian-distributed random variable with a standard deviation of about one fourth of the average (Dissanayake et al., 1997), so it is a tight distribution.

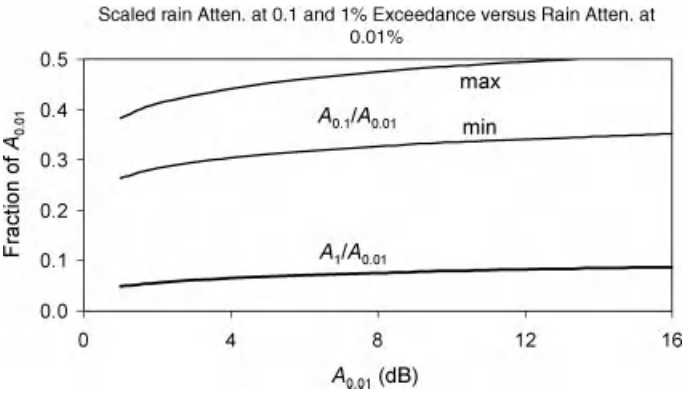


FIGURE 11.4 Scaled rain attenuation at exceedances 0.1 and 1% versus rain attenuation at 0.01%, for GEO elevation at least 20°.

TABLE 11.5 Gaseous Attenuation on Link with ITU-R's Mean Annual Global Reference Atmosphere

Carrier Frequency (GHz)	Atten. (dB) for 90° Elev.	Atten. (dB) for 30° Elev.	Atten. (dB) for 20° Elev.
12	0.06	0.12	0.18
20	0.27	0.54	0.79
30	0.25	0.50	0.73

The only time it may make sense to use something besides a constant loss for each location of interest is for a dry climate, low elevation angle, and high exceedance (Feldhake, 1997).

The ITU-R gives a rather lengthy way (in Annex 2 of Recommendation, ITU-R, P.676-7) to compute gaseous attenuation from atmospheric pressure, temperature, and water-vapor information at the ground station. There are two slightly different versions of the computation, one using surface water-vapor density and the other using integrated water-vapor content as a function of exceedance. The ITU-R provides several reference standard atmospheres for which it gives the three parameters of the first version: mean global, low-latitude annual, mid-latitude summer and winter, and high-latitude summer and winter (Recommendation, ITU-R, P.835-4). The ITU-R also provides maps of the integrated water-vapor content for exceedances from 1 to 20% (Recommendation, ITU-R, P.836-3). Annex 2 of Recommendation, ITU-R, P.676-7 plots attenuation versus frequency on a zenith path for the mean annual global reference atmosphere. (Warning: the plot was wrong in version 6 of that document and was corrected in version 7.) Values from that plot are given in Table 11.5, along with values for 30 and 20° elevation.

The attenuation from clouds, which is due to the water vapor in them, can be significant at frequencies well above 10 GHz (Recommendation, ITU-R, P.840-4) at high exceedances, when there is no rain, but not at the low exceedances typically necessary for a communications link (Dissanayake et al., 1997; Recommendation, ITU-R, P.840-4).

Another atmospheric effect we can neglect is tropospheric scintillation, which can be important in some scenarios for frequencies above 10 GHz. The scintillation is caused primarily by clear-air turbulence in the atmosphere's horizontally stratified layers that have different refractive indexes. It occurs when there are high humidity gradients and temperature inversion layers. Amplitude scintillation can be significant for elevation angles below 10°, but this is an unusually low elevation angle for satellite communications. It is primarily a clear-air problem (Ippolito, 2008), that is, not a problem at the low exceedances typically necessary for a communications link.

A way to combine the atmospheric attenuations is needed, for a link operating at frequencies above about 18 GHz and especially if it has a low elevation angle or atmospheric margin. The ITU-R provides this (Recommendation, ITU-R,

P.618-10). The ITU-R found that this formula matched the multiyear measurements in their database to an rms of about 25%, at all exceedance levels. The model is available as Excel macros on the Internet site logiciel.cnes.fr/PROPA/en/logiciel.htm. When clouds and tropospheric scintillation can be neglected, the combined attenuation is the sum of the rain attenuation at the exceedance of interest and the gas attenuation at 1% exceedance if known from local data, otherwise the mean gas attenuation.

11.3.7 Loss and Variation from Antenna Polarization Mismatch

For the receiving antenna to be able to receive the entire radiation incident on it from a transmitting antenna, the radiation and the receiving antenna must have the same polarization and the same alignment over the 180° range of possibilities. When this is not the case, there is **polarization mismatch loss**. For a GEO with fixed ground station, the loss would be constant but in other situations it will most likely vary. The only time this loss is likely to be rather large is when one of the antennas is either a phased-array antenna or has a phased-array feed and the beam is scanned (Section 3.4.2 or 3.5.3, respectively).

The relative amount of signal power received compared to the power available, due to polarization mismatch, is the following (Howard, 1975):

$$\rho = \frac{1}{2}\frac{(1 \pm r_1 r_2)^2 + (r_1 \pm r_2)^2 + (1 - r_1^2)(1 - r_2^2)\cos 2\theta}{(1 + r_1^2)(1 + r_2^2)}$$

where $r_1 =$ axial ratio of elliptically polarized wave, $r_2 =$ axial ratio of elliptically polarized receiving antenna, $\theta =$ angle between direction of maximum amplitude in

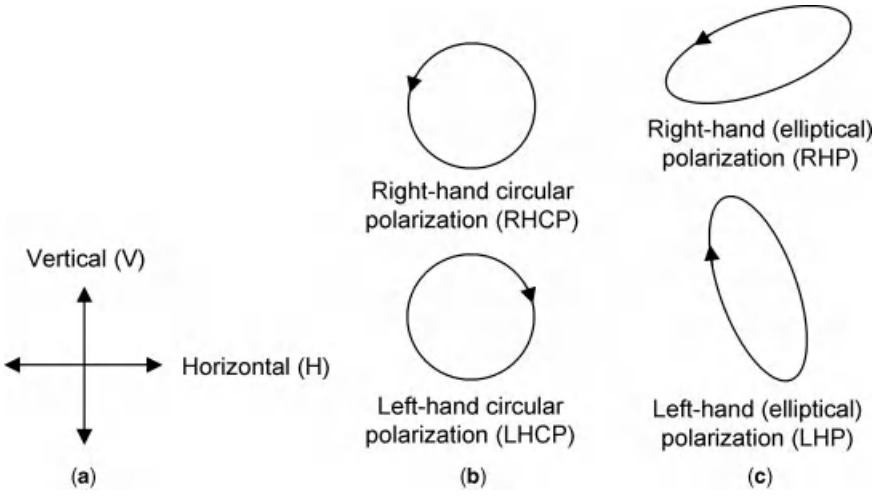


FIGURE 11.5 Pairs of orthogonal antenna polarizations: (a) linear, (b) circular, and (c) (nonstandard) elliptical.

incident wave and direction of maximum amplitude of elliptically polarized receive antenna, and + sign is to be used if and only if both the receiving and transmitting antennas have the same "hand" of polarization, otherwise − sign. Both r_1 and r_2 are between 0 and 1, equaling 0 for linear polarization and 1 for circular polarization (Section 3.2.3.4). When $r_1 = 1$, $\rho = (1 + r_2)^2/(2(1 + r_2^2))$. For polarization that is not linear, typically not r_1 and r_2 but $-20 \log r_1$ and $-20 \log r_2$ (in dB) are given. Normally, for example, in a link budget, the polarization mismatch loss would be reported as $10 \log \rho$. If the two polarizations are orthogonal ($r_1 = r_2$, the "hands" are different, and $\cos 2\theta = -1$), no radiation will be received. Examples of pairs of orthogonal polarizations are shown in Figure 11.5. If a CP antenna receives LP, it receives half the power, and vice versa.

11.4 NOISE LEVEL ON LINK

The system noise temperature of the link will vary for several reasons. An overview of the variations is given in Table 11.6. Each end of a link must keep the changes it causes within a certain range, and the receiving end must be able to accommodate the changes outside its control.

As we have seen, the system noise temperature of the payload or ground station is the sum of the antenna-noise temperature and the electronics-noise temperature. The second and third have been discussed elsewhere in this book, and here we describe the antenna-noise temperature T_a. It is a property of the link. T_a is not actually due to the antenna but to the background radiation that the antenna receives along with the signal, averaged over the antenna pattern. The ITU-R characterizes antenna-noise temperature for both links (Recommendation, ITU-R, P.372-9). The general calculation of antenna-noise temperature is the convolution of the antenna pattern with the "brightness temperature" of the background, where the brightness temperature is frequency-dependent. The ITU-R provides plots and maps and tells

TABLE 11.6 Variations in System Noise Temperature

Variation Type	Source	Section
Variation in payload antenna temperature	Temperature of earth below a moving transmitter varies	11.4
Variation in ground antenna temperature	Sun or moon in antenna beam, heavy rain	11.4
Jump up in payload front-end noise figure	Switch to redundant receiver (more switches before LNA)	4.6.2
Jump up in ground receiver front-end noise figure	Switch to redundant receiver (more switches before LNA)	N/A
Jump in payload back-end noise figure (preamp in ALC mode)	Preamp ALC changes attenuation in reaction to change in input signal level	7.12.1
Jump in ground receiver back-end noise figure	ALC changes attenuation in reaction to change in input signal level	N/A

how to compute the brightness temperature. Following is general information from that document.

For a payload receiving a signal with the earth as background, the brightness temperature is due to a combination of atmospheric radiation reflected from the earth and radiation emitted by the earth. A water surface has a brightness temperature from zenith of about 95 K at 1.4 GHz, 110 at 10 GHz, and 175 at 37.5 GHz, depending on the water surface temperature, the salinity, and the roughness. The land is brighter than water. The drier and the rougher the land is, the brighter it is. At 1.43 GHz, the temperature from zenith ranges from about 180 to 280 K. For a payload receiving a signal from another satellite with cold space as the background, at frequencies above 1 GHz the brightness temperature is only a few degrees except in the direction of the sun, which has a very high temperature, and the moon, which varies from about 140 K at new moon to 280 K at full moon.

For a ground station receiving a signal from a payload, the sky is the background. In the range of frequencies from 1 to 30 GHz, about 23 GHz has the maximum brightness temperature of the clear-sky atmosphere. There, for a zenith path the temperature is about 28 K and for a path of 30° elevation it is about 53 K. For frequencies above 20 GHz, heavy rain increases the sky temperature, up to a limit of an extra 290 K in a monsoon (Morgan and Gordon, 1989). The brightness temperature of the sun and the moon are the same as for a payload receiving a signal from another satellite with cold space as the background. The link will be unavailable once a day for several days around the time when the sun is exactly behind the spacecraft (Maral and Bousquet, 2002).

11.5 INTERFERENCE ON LINK

11.5.1 Introduction

Interference can be divided into two categories based on source: self-interference and interference from other systems. Self-interference is a satellite system interfering into itself. By "satellite system" we mean (although other definitions are possible) one satellite or collocated satellites (in the same orbital slot, with the same owner) and its/their users; a transmitter to (or receiver from) a different satellite but collocated with our transmitter (or receiver, respectively) does not count as self-interference. Interference from other systems can come from other satellite systems and, in some frequency bands, from terrestrial communications systems. Interference can occur on the uplink and on the downlink.

When interference into a satellite system is a possibility, whether from self or not, the system design must create discrimination against the interference in at least one of the following three domains or "dimensions," as illustrated in Figure 11.6:

- *Frequency*, that is, the interfering signal is in a different part of the frequency band from the signal of interest. Specifically, at the output of the signal of interest's detection filter, the power of the sampled signal of interest is greater than the power of the interfering signal

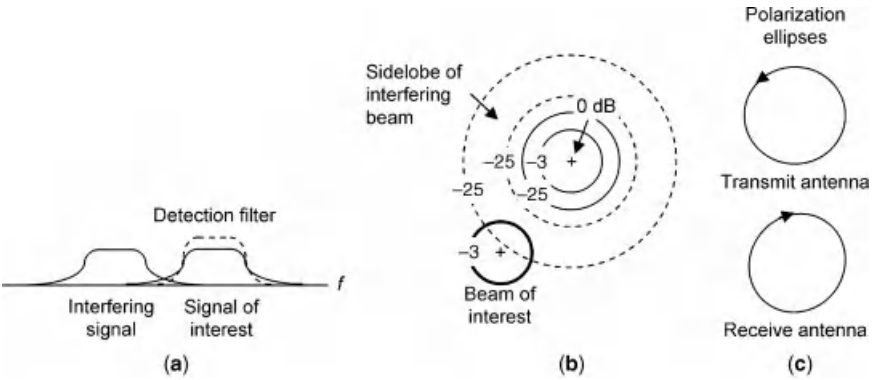


FIGURE 11.6 Three dimensions of discrimination against interfering signal: (a) frequency, (b) antenna gain, and (c) polarization.

- *Antenna pattern*, that is, the signal of interest is being transmitted and received on the (high-gain part of the) mainlobe of both antenna patterns but the same cannot be said for the interfering signal
- *Polarization*, that is, the interfering signal nominally has the orthogonal polarization to that of the signal of interest.

Increasing the power of the signal of interest would help but is usually not an option. When the signal of interest is weaker than the interfering signal, it is harder to bring E_s/I_0 into an acceptable range, where $E_s/I_0 = (C/I_0)/R_s$ and R_s is modulation-symbol rate (Section 10.6). For frequencies above 10 GHz, attenuation from rain on the link can be significant, which suppresses the signal of interest relative to its interferers. The difference between rain attenuation on two links with the same ground station is believed unlikely to tip the balance between acceptable and unacceptable interference (Recommendation, ITU-R, P.619-1).

Probably a composite E_s/I_0, sampled out of the detection filter, of at least 20 dB is adequate for link closure. Usually if there is good discrimination in at least two dimensions, E_s/I_0 is much better than this. In the following three subsections, we look at cases where there is no discrimination in two dimensions and good but not great discrimination in the third.

11.5.2 Adjacent-Channel Interference

Adjacent-channel interference (ACI) is frequency-domain interference. We assume that the antenna gain and polarization for the interfering signal are the same as for the signal of interest. Possible causes of significant ACI are (1) multiplexing of channel frequency bands that does not sharply enough filter out adjacent channel(s) and (2) spectrum spreading of the interfering signal by its HPA (Section 8.4.2.2). ACI is self-interference and can arise in the transmitting ground station, the payload, and/or in the receiving ground station. Adjacent channels are normally

separated by a guard band about 10% as wide as the channel spacing (Section 4.5.3.4). If the signals on the two channels have different spectral bandwidths but their power spectral densities are about the same, the wider signal is more likely to interfere into the narrower one than the other way around.

11.5.3 Sidelobe Interference

"Sidelobe interference" is antenna pattern-domain interference. This interference comes from an interfering signal that is either transmitted on a sidelobe and received on a mainlobe or vice versa. We assume that the frequency band is the same for the interfering signal as for the signal of interest, and we make the worst-case assumption that the polarization is the same. Figure 11.7 shows the scenarios that can lead to sidelobe interference, for the uplink and downlink. Half of them are self-interference. Only interference scenarios from a satellite system are shown in Figure 11.7; scenarios from a terrestrial system can exist but are not shown. The geometry of each uplink scenario is basically the same as that of a particular downlink scenario, as shown in Figure 11.7. Uplink scenarios (a) and (b) involve reception by a sidelobe, while downlink scenarios (a) and (b) involve transmission by a sidelobe. Scenarios (c) and (d) are opposite of this. Scenario (d) involves two different ground stations that are close enough to be in the same satellite beam.

Actually, a sidelobe need not be involved, since the low-gain outer edge of a mainlobe could cause a similar situation. For a coverage area tiled with three-frequency-reuse spot beams (Section 3.2.1), in the spot of interest the interfering

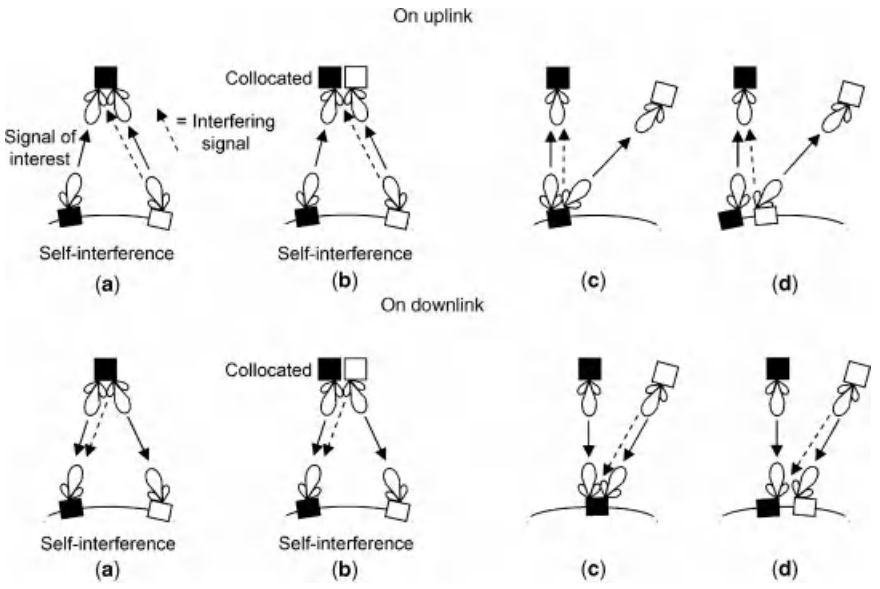


FIGURE 11.7 Scenarios of sidelobe interference from a satellite system, on uplink and downlink.

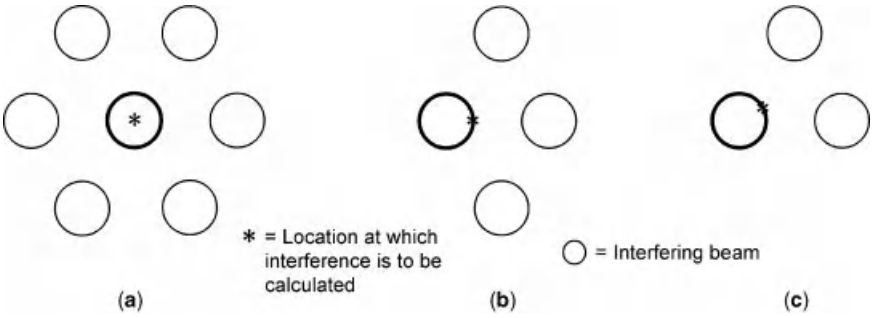


FIGURE 11.8 Significant interfering beams for various locations in spot beam.

spot beam is probably transmitting/receiving on the side of its mainlobe, while for a coverage area tiled with four-frequency-reuse spot beams, the interfering spot beam is probably transmitting/receiving in its first sidelobe (the latter depicted in Figure 11.6b) (Rao, 2003). Figure 11.8 shows probably the most significant interfering spot beams for various locations of interest relative to surrounding spot beams (the closer beams not shown are on different channel bands). When there are a few interfering spot beams, the total interference can be significant even though the interference from one beam is not (see Rao (2003) for example).

11.5.4 Cross-Polarized Interference

Interference from cross-polarized (cross-pol) signals is polarization-domain interference. We assume that the frequency band and antenna gain for the interfering signal are the same as for the signal of interest. Then this type of interference can arise on either link when two orthogonal polarizations are being transmitted. The interfering signal arriving at the receiver must have significant content in the polarization of the signal-of-interest's receive beam. The transmit and receive antennas may provide this directly by not being polarized exactly the same, or heavy rain may corrupt the interfering signal's polarization. Figure 11.9 shows the scenarios for cross-pol interference for uplink and downlink. All of them are self-interference. The geometry of each uplink scenario is basically the same as that of a downlink scenario, as shown in Figure 11.9. Scenarios (a) and (b) involve one satellite, while (c) and (d) involve collocated satellites. Scenarios (a) and (c) involve two ground stations close enough to be in the same satellite beam, while (b) and (d) involve one ground station.

Heavy rain can cause cross-pol interference. Rain depolarization is significant for 10 GHz and higher frequencies, that is, the same frequencies at which rain attenuation is significant. Heavy rain like this degrades E_s/I_0 in two ways, one by increasing the interference and the other by depressing the signal power. The amount of depolarization of a signal is usually characterized in terms of the parameter **cross-polarization discrimination (XPD)**. Suppose that a signal path goes through rain. XPD is the ratio in dB at input to the receive antenna of the amount of

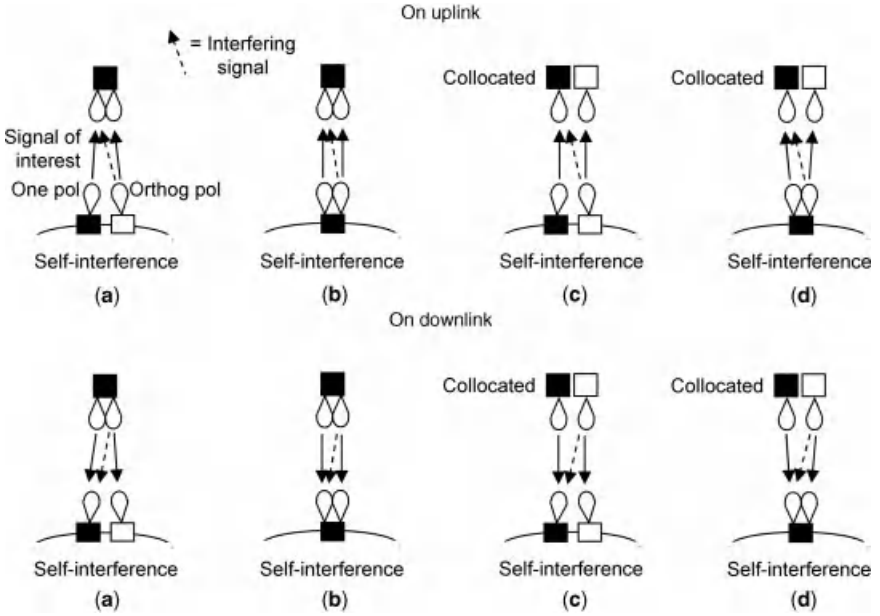


FIGURE 11.9 Cross-pol interference scenarios on uplink and downlink.

power in the transmitted polarization to the amount in the orthogonal polarization. XPD in dB is linearly related to the logarithm of the copolarized rain attenuation in dB on a path (Recommendation, ITU-R, P.618-10). XPD is up to 15 dB greater (more benign) for linearly polarized signals than for circularly, depending on the polarization tilt angle (Section 11.3.5). For the horizontal polarization, it is 15 dB, while for the vertical polarization it is 15 dB at the subsatellite point, decreasing to 0 dB away from the subsatellite point, and reaching 15 dB again at the horizon (the earth's limb). In the next paragraph, we refer to circular polarization but we could just as well be speaking of vertical polarization toward that circle of 0 dB difference. Raising the signal power does not improve XPD.

Besides depending on polarization tilt angle, the coefficients for the linear relationship of XPD to the logarithm of attenuation in dB depend on frequency, elevation angle to the satellite, and almost negligibly on the standard deviation of the raindrop "canting-angle" distribution (Recommendation, ITU-R, P.618-10). As raindrops fall, they are flattened by air resistance. Their canting angle is 0° when they fall straight down and greater than 0° when they make an angle with the horizontal due to wind. The canting-angle standard deviation is about 10° for rain attenuation (and XPD) at the 0.01% exceedance level (Ippolito, 2008). Figure 11.10 plots XPD as a function of rain attenuation and elevation angle for a circularly polarized signal at 12, 15, 20, and 30 GHz.

It has been shown that for a dual-polarized system, the sensitivity to rain-caused cross-pol reduces as the required E_s/N_0 goes down (Vasseur, 2000). We can conclude three things: (1) error-control coding reduces the sensitivity; (2) higher-order

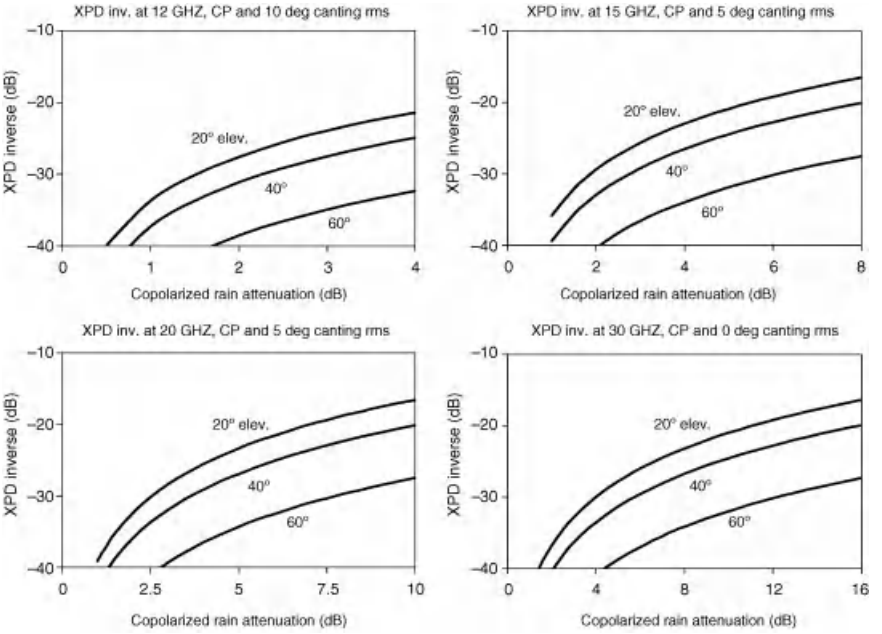


FIGURE 11.10 Cross-polarization discrimination versus copolarized rain attenuation at 12, 15, 20, and 30 GHz, for circularly polarized signal.

modulation increases the sensitivity; and (3) the ground stations must have excellent polarization isolation.

11.6 LINK BUDGET

We now put together what we have discussed so far in this chapter into a **link budget**, which is a tallying up of the transmitted power, power losses, and noise and interference contributions to a bottom-line $E_s/(N_0 + I_0)$ and then comparison of this with the $E_s/(N_0 + I_0)$ necessary for adequate performance. In most cases, interference is well modeled as an additional noise-type term (see Section 13.2.4 for a discussion of this), which is why N_0 and I_0 are usually summed as they are here. The correct sign is put on entries so that they can all be summed, that is, if an item is called a "loss" it should still be shown negative. In this particular sample budget, the term called "implementation loss" is a catch-all for less-than-ideal behavior of the system, which is not well characterized, and includes payload imperfections, ground station imperfections, and payload-ground interaction imperfections. The term "*C*/*I* before discrims applied" is the ratio of power input to the antenna terminal for *C* relative to that for *I*, before any discrimination is considered. The choice to have no discrimination in the frequency domain and 10 dB in each of the antenna-pattern and polarization domains is random and just for illustration. The link budget we present in

Term		Unit	Value
GEO transmit			
TWTA pout		dBW	16.8
Post-TWTA loss		dB	−1.7
Redundancy switch		dB	−0.3
Antenna gain		dB	55.3
Pointing loss		dB	−0.3
	EIRP	dBW	69.8
Propagation			
Free space loss		dB	−210.3
Gas attenuation		dB	−0.5
Rain atten (0.01%)		dB	−5.5
Polarization mismatch		dB	−0.5
Pointing loss		dB	−0.3
Antenna gain		dB	69.0
	C	dBW	−78.3
Noise			
System noise temperature, inverse		dB(1/K)	−28.9
Boltzmann's constant, inverse		dB(K-Hz/W)	228.6
	C/N_0	dBHz	121.4
Interference			
C/I before discrims applied		dB	0.0
Frequency discrimination		dB	0.0
Antenna gain discrimination		dB	10.0
Polarization discrimination		dB	10.0
Symbol rate		dBHz	99.8
	C/I_0	dBHz	119.8
Combined			
	$C/(N_0+I_0)$	dBHz	117.5
Symbol rate, inverse		dB(1/Hz)	−99.8
	$E_s/(N_0+I_0)$	dB	17.7
Margin			
Implementation loss		dB	−1.5
Risk margin		dB	−2.0
Required $E_s/(N_0+I_0)$, inverse		dB	−13.0
	Excess margin	dB	1.2

FIGURE 11.11 Sample downlink budget.

Figure 11.11 is a simplified one. Ordinarily, there would also be explanatory entries that are not part of the summation but are used in calculation of entries that are summed. Also, one may want to make maximum and minimum signal-level cases. Uncertainties or variations could also be included (Section 8.3.2).

REFERENCES

Allnutt JA (1989). *Satellite-to-Ground Radiowave Propagation*. London: Peter Peregrinus Ltd.

Boonchuk T, Hemmakorn N, Supnithi P, Iida M, Tanaka K, Igarashi K, and Moriya Y (2005). Rain attenuation of satellite link in Ku-band at Bangkok. *Proceedings of International*

Conference on Information, Communications and Signal Processing; Dec. 6–9; 1093–1096.

Crane RK (1993). Estimating risk for earth-satellite attenuation prediction. *IEEE Proceedings;* 81 (June); 905–913.

Crane RK (1996). *Electromagnetic Wave Propagation through Rain*. New York: John Wiley & Sons, Inc.

Crane RK (2002). Analysis of the effects of water on the ACTS propagation terminal antenna. *IEEE Transactions on Antennas and Propagation;* 50 (Jul.); 954–965.

Crane RK (2003a). A local model for the prediction of rain-rate statistics for rain-attenuation models. *IEEE Transactions on Antennas and Propagation;* 51 (Sept.); 2260–2273.

Crane RK (2003b). *Propagation Handbook for Wireless Communication System Design*. Boca Raton, FL: CRC Press.

Crane RK and Dissanayake AW (1997). ACTS propagation experiment: attenuation distribution observations and prediction model comparisons. *IEEE Proceedings;* 85 (June); 879–892.

Davies K and Smith EK (2002). Ionospheric effects on satellite land mobile systems. *IEEE Antennas and Propagation Magazine;* 44 (Dec.); 24–31.

Dissanayake A, Allnutt J, and Haidara F (1997). A prediction model that combines rain attenuation and other propagation impairments along Earth-satellite path. *IEEE Transactions on Antennas and Propagation;* 45; 1546–1558.

Feldhake G (1997). Estimating the attenuation due to combined atmospheric effects on modern earth-space paths. *IEEE Antennas and Propagation Magazine;* 39; 26–34.

Green HE (2004). Propagation impairment on Ka-band SATCOM links in tropical and equatorial regions. *IEEE Antennas and Propagation Magazine;* 46 (Apr.); 31–45.

Howard W. Sams & Co, Inc. (1975). *Reference Data for Radio Engineers*. Indianapolis, IA: Howard W. Sams & Co, Inc.

Ippolito LJ (2008). *Satellite Communications Systems Engineering: Atmospheric Effects, Satellite Link Design, and System Performance*. UK: John Wiley & Sons Ltd.

Mandeep JS and Allnutt JE (2007). Rain attenuation predictions at Ku-band in South East Asia countries. *Progress in Electromagnetics Research;* 76; 65–74.

Maral G and Bousquet M (2002). *Satellite Communications Systems*, 4th ed. Chichester, England: John Wiley & Sons Ltd.

Morgan WL and Gordon GD (1989). *Communications Satellite Handbook*. New York: John Wiley & Sons, Inc.

NASA, Goddard Space Flight Center (2010). TRMM, Tropical Rainfall Measuring Mission. On trmm.gsfc.nasa.gov. Accessed Sept. 10, 2010.

Ramachandran V and Kumar V (2006). Modified rain attenuation model for tropical regions for Ku-band signals. *International Journal of Satellite Communications and Networking;* 25; 53-67.

Rao SK (2003). Parametric design and analysis of multiple-beam reflector antennas for satellite communications. *IEEE Antennas and Propagation Magazine;* 45 (Aug.); 26–34.

Recommendation, ITU-R, P.372-9 (2007). Radio noise. Geneva: International Telecommunication Union, Radio Communication Section.

Recommendation, ITU-R, P.531-10 (2009). Ionospheric propagation data and prediction methods required for the design of satellite services and systems. Geneva: ITU-R.

Recommendation, ITU-R, P.618-10 (2009). Propagation data and prediction methods required for the design of Earth-space telecommunication systems. Geneva: ITU-R.

Recommendation, ITU-R, P.619-1 (1992). Propagation data required for the evaluation of interference between stations in space and those on the surface of the earth. Geneva: ITU-R.

Recommendation, ITU-R, P.676-7 (2007). Attenuation by atmospheric gases. Geneva: ITU-R.

Recommendation, ITU-R, P.835-4 (2005). Reference standard atmospheres. Geneva: ITU-R.

Recommendation, ITU-R, P.836-3 (2001). Water vapor: surface density and total columnar content. Geneva: ITU-R.

Recommendation, ITU-R, P.837-5 (2007). Characteristics of precipitation for propagation modelling. Geneva: ITU-R.

Recommendation, ITU-R, P.840-4 (2009). Attenuation due to clouds and fog. Geneva: ITU-R.

Vasseur H (2000). Degradation of availability performance in dual-polarized satellite communications systems. *IEEE Transactions on Communications;* 48; 465–472.

CHAPTER 12

PROBABILISTIC TREATMENT OF MULTIBEAM DOWNLINKS

12.1 INTRODUCTION

This chapter presents a link-level analysis of a multibeam payload that shows lower EIRPs to be necessary than a traditional, overly conservative analysis would dictate. The new analysis treats variations via probability theory. (Here, the word "variation" is used instead of "uncertainty" as in Chapter 8, even though the definition of variation is the same as the definition of uncertainty, because "variation" is the term used in a payload specification.) Even relatively more EIRP can be saved if the downlinks are subject to rain loss and the traditional EIRP requirements are replaced with a link-availability requirement. Recall that the downlink is available if the link "closes," that is, if the signal arriving at the ground-station receiver meets a minimum SNR requirement.

In a traditional analysis of this kind, worst-case value is piled upon worst-case value and held against C (=EIRP) and used to build up I_{self} (=self-interference—see Section 11.5). This characterizes an extreme situation that occurs with miniscule probability. A probabilistic characterization more fully and accurately portrays the on-orbit payload performance. If the links are subject to rain loss, traditionally a rain margin is separated out from a payload-variation margin and applied one-to-one to increase the specified EIRPs. In reality, though, most of the time the payload's performance is fairly close to nominal, so for the rare times when the rain attenuation is uncommonly large, the chance is high that the payload is performing near nominal.

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

In this chapter, we deal with the following determinants of the power that reaches the ground-station:

- Varying performance of repeater, that is, payload minus antennas
- Varying performance of payload antennas
  - Antenna-gain inaccuracy
  - Additional gain inaccuracy in reflector antenna from diurnal pattern distortion
  - Antenna-pointing error
- Atmosphere, for carrier frequencies above about 10 GHz (see Section 11.3.6).

The I_{self} that matters most in C/I_{self} variation is for the interferer that creates the smallest C/I_{self} for the particular C of interest, at the coverage-area location under consideration. C/I_{self} includes the discrimination that C has against the self-interference, but the C/I_{self} variation does not. At any randomly chosen location, there is usually just one significant interferer, if any. There may well be isolated points in the coverage area at which more than one interferer is significant.

The rest of the regular sections of this chapter are as follows:

- *Section 12.2*: Multibeam-downlink payload specifications
- *Section 12.3*: Probabilistic characterization (i.e., standard deviation and nominal value) of the repeater contribution to C and C/I_{self}
- *Section 12.4*: Probabilistic characterization of the payload antenna's contribution to C and C/I_{self} and combining it with repeater contribution
- *Section 12.5*: Probabilistic characterization of $C/(I_{\text{self}} + I_{\text{other}} + N)$ from payload sources, where I_{other} and N are fixed. I_{other} is interference from other satellites and terrestrial sources, and N is noise from the receiving ground station
- *Section 12.6*: Link availability from statistical atmospherics-attenuation model combined with payload-caused variation
- *Section 12.7*: Iterative optimization of multibeam-downlink payload specified on link availability
- *Appendix 12.A*
  - *Section 12.A.1*: Details of iterative technique for optimizing multibeam payload with link-availability specification
  - *Section 12.A.2*: Probability density function (pdf) (Section 8.A.2) of diurnal variation in delta of east and west panel illumination.

12.2 MULTIBEAM-DOWNLINK PAYLOAD SPECIFICATIONS

There are different kinds of multibeam-downlink payloads and ways to specify them. The basic specification is either EIRP or link availability. For carrier frequencies below 10 GHz, it has to be EIRP since there is no significant atmospheric loss

(for uncommon exceptions see Section 11.3.5). For frequencies above 10 GHz, it can be either EIRP or link availability. Whichever it is, other specifications follow along. First we discuss the kinds of specifications and then the set of locations on which they must hold.

If the basic specifications are a set of minimum EIRPs then these EIRPs compensate rain attenuation, if any. Other specifications are maximum C variation, minimum C/I_{self}, and maximum C/I_{self} variation. (With the probabilistic approach given in this chapter, it should be easy to meet the variation specifications.)

If the basic specification is a minimum link availability then a particular statistical rain model is also specified. Also specified is the minimum threshold of atmosphere-attenuated $C/(I_{\text{self}} + I_{\text{other}} + N)$ for the ground-station receiver to perform acceptably. The threshold equals the value necessary to the receiver if the only signal impairment were additive white Gaussian noise, augmented to make up for the less-than-ideal implementations of the ground stations and the payload and for their interactions. I_{other} and N are specified, both constant and both referred to the same point as C, namely at the payload-antenna output. Atmosphere-attenuated $C/(I_{\text{self}} + I_{\text{other}} + N)$ must exceed the threshold for a percentage of the time at least equal to the specified minimum availability, long-term.

Link availability instead of EIRP is recommended as the basic specification for a multibeam-downlink payload with carrier frequencies 10 GHz or higher. Availability is fundamentally what counts. Making a satellite create and transmit more power than necessary carries a cost penalty.

In either case, there is also a set of maximum EIRP specifications, to ensure that the **power flux density** reaching the earth (see Figure 12.1) meets regulations, within and in some cases outside of the coverage area.

Now, on what set of locations do the specifications apply? Multiple beams may be all of the same size (as seen from the payload) in a hexagonal grid or of different sizes and in an irregular pattern. For carrier frequencies below 10 GHz, the specifications can apply to "every square inch" of the coverage area because it is easy to see what the possible worst locations are. They are the same in every beam relative to its center. It is comparatively easy to check those few locations and make adjustments in payload design. However, for carrier frequencies above 10 GHz, the

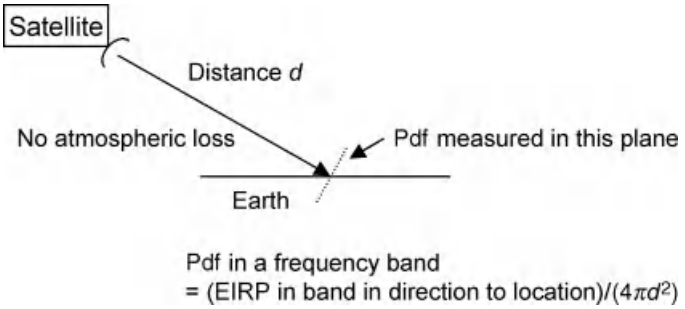


FIGURE 12.1 Definition of power flux density reaching location on earth.

specifications can only apply to a (large number of) discrete locations. The beam pattern is necessarily irregular at the least in the sense of varying power levels in the beams to compensate atmospheric attenuation. It is usually not possible to look at an irregular beam pattern and the irregular rain map and see where the tricky locations are. It is humanly impossible to check specifications on an infinite number of locations. All one can do is check a large number of locations.

12.3 REPEATER-CAUSED VARIATION OF C AND C/I_{self} AND NOMINAL VALUE

12.3.1 Introduction

The following analysis of the repeater-caused variation in C and C/I_{self} is based on the probabilistic treatment of uncertainty in payload budgets of Section 8.3.2. The analysis is made at a point in time at which units are being built but the TWTAs are complete, but with small changes the analysis could be adapted to any other time in the payload build (Section 8.3.2.2). We also find the nominal value of C and C/I_{self} relative to their beginning-of-life (BOL) values (Section 8.3.2.9). For one interfering signal I_{self} we also analyze the C/I_{self} variation, which we need in the later analysis of the $C/(I_{\text{self}} + I_{\text{other}} + N)$ variation.

The type of "lifetime" requirement for the payload specifications matters (Section 8.3.2.5). Let us define δ as in Section 8.3.2.9, namely as the expected change in average performance from BOL to middle of life, in dB. For the "average over life" lifetime requirement, the variation is averaged over life and the nominal value is δ relative to BOL. For "at EOL" both the variation and the nominal are for EOL, so the nominal is 2δ relative to BOL. Results are primarily shown for "average over life" but adaptations for "at EOL" are stated. We will find that the repeater has a nonzero contribution to δ but the payload antenna contributes zero.

The repeater is assumed to have the block diagram presented in Figure 1.4 and the layout in the spacecraft shown in Figure 2.4a. The CAMPs are operated in automatic-level-control (ALC) mode. The C variation due to the repeater is the following sum:

$$\begin{aligned} C \text{ variation due to repeater in dB} = {} & \text{Variation in TWTA } P_{\text{out}} \\ & + P_{\text{out}} \text{ measurement error} \\ & + \text{Variation in repeater loss past TWTA, all in dB} \end{aligned}$$

That is, the variation comes only from the CAMP (since it drives the TWTA), the TWTA, the OMUX, and the rest of the post-TWTA hardware (often just waveguide) except the antenna. The units on the signal path before the CAMP do not matter because the CAMP compensates input signal-level variation since it is in ALC mode. The variation in I_{self} is similar to that for C but from different payload equipment.

The orientation of the GEO satellite in space relative to the earth and the sun at various times of year is illustrated in Figures 2.7–2.9. The variation of C has a term

TABLE 12.1 Independent Components of Repeater-Caused C Variation, as Analyzed

Constant but unknown offsets
Diurnal variations
Lifetime drifts unrelated to temperature

that is constant but unknown over life, a diurnal swing (different at BOL and EOL), a seasonal variation, and a drift of the year's average value over the years of life. We make the pessimistic assumption that every day has the EOL, worst-day diurnal spread. The seasonal variation is minor, so we make the slightly pessimistic simplification of only treating the worst day of the year. The constant but unknown offset and the variations on the remaining time scales are now independent and are listed in Table 12.1.

In the assumed layout of the payload in the spacecraft, the CAMPs and TWTAs are on the north and south panels and the OMUXes are on the east and west panels. There is temperature variation of the spacecraft mounting plates to which the repeater's unit baseplates are attached. The north and south panels have heat pipe systems which are connected, so the temperatures of units on the two panels are always nearly the same. The east and west panels have their own heat pipe systems, separate from each other and the heat pipes of the north and south. There will be more temperature difference at any one time among the OMUXes than among the CAMPs or among the TWTAs. The nominal temperatures of these units are shown in Table 12.2.

For the C variation due to temperature change we are interested in the temperature variations of the individual units in the signal path, while for the C/I_{self} variation we are interested in the variation of the temperature delta between two units of the same kind. Only one unit in any signal path, the OMUX, has a diurnal temperature swing. We assume that the drive levels into the TWTAs stay very nearly the same over life. Some examples of realistic numbers for single units are given in Table 12.3. Between two OMUXes, if they are on the same panel the difference is zero (i.e., they are completely positively correlated) and if they are on different panels the limits are twice as great as for single units (i.e., they are completely negatively correlated).

TABLE 12.2 Nominal Temperatures of Units (Example)

Unit	Nominal Temperature Relative to BOL Average (°C)
CAMP	+35
TWTA	+35
OMUX on east panel	+27
OMUX on west panel	+27

TABLE 12.3 Temperature Variation of Single Units (Example)

Unit	EOL Diurnal Temperature Swing (°C)
CAMP	0
TWTA	0
OMUX on east panel	±27
OMUX on west panel	±27

The analysis flow is as follows. Sections 12.3.2–12.3.6 break down the C variation and the C/I_{self} variation due to the repeater, each into three independent components. Section 12.3.7 lists the units' contributions to the variations, representing them by random numbers. Section 12.3.8 combines the units' contributions into the composite variations and computes the standard deviations. Also derived is the parameter δ. A numerical example is carried throughout. The flow will not be broken, and may even be enhanced, if the reader skips the details of the breakdown on the first reading and goes immediately to Section 12.3.8. In the following tables and equations, "negligible" means less than 0.05 dB.

12.3.2 Variation Contributions from CAMP

The CAMP preamplifies the signal to the level desired for input to the HPA. Recall that in our example all CAMPs are operated in ALC mode.

12.3.2.1 CAMP Settability Resolution The CAMP's output power is settable to steps with a resolution of 0.5 dB, so without particular information on the settability error we assume it is uniformly distributed between −0.25 and +0.25 dB. Settability resolution affects only the signals with unsaturated TWTAs, where it leads to an error in the TWTA's P_{out}. Half the values of P_{out} variation due to settability resolution are beneficial to C and half are detrimental.

Table 12.4 shows examples of the variation in P_{out} due to one CAMP's settability resolution. For two signals, it is likely that either the settabilities are the same, since the CAMPs may come from the same build, or independent. Independence is worse so we assume that. Then the variation in the ratio of two signals' P_{out} due to their CAMPs' settability resolutions is the sum of the two individual, independent random numbers.

TABLE 12.4 C Variation due to CAMP's Settability Resolution (Example)

	Constant but Unknown Offset
	In general form of a random number uniformly distributed over $-Q_1$ to $+Q_1$ where Q_1 is
TWTA at saturation	0 dB
TWTA at small-signal	0.25 dB

TABLE 12.5 C Variation due to CAMP's ALC-Mode Maximum-Power Variation (Example)

	Lifetime Drift Unrelated to Temperature
	At EOL, in general form of a random number uniformly distributed over $-L_1$ to $+L_1$ where L_1 is
TWTA at saturation	0.1 dB
TWTA at small-signal	0.4 dB

12.3.2.2 CAMP Temperature Compensation The CAMP attempts to compensate its ALC-mode output power for the TWTA's P_{out} variation with temperature but it does not compensate exactly. Over the operating temperature range from -5 to 60°C, CAMP output power rises linearly by 0.5 dB. This temperature compensation is taken into account in the section below on TWTA variation over temperature.

12.3.2.3 CAMP Maximum-Power Variation due to Aging and Radiation We assume that the CAMP's ALC-mode maximum power has an unknowable linear drift over life that may be either positive or negative. Half the values of drift are beneficial to C and half are detrimental. The size of the drift at EOL is uniformly distributed between zero and the maximum positive EOL drift.

Table 12.5 shows examples of the lifetime drift in P_{out} due to one CAMP's ALC-mode maximum-power drift. For two signals, it is likely that either the drifts are in the same direction or are independent. Independence is worse so we assume that. The pdf of one and the pdf of the sum of two independent drifts averaged over life are given in Appendix 8.A.7.

12.3.3 Variation and Nominal-Value Contributions from TWTA

12.3.3.1 P_{out} Nominal Value from Temperature For most TWTAs in a flight set with a constant drive level, P_{out} goes down a little as the TWTA's temperature rises, due to the TWTA's loss increasing. A TWTA will experience no diurnal variation (see Table 12.3), so there is also no diurnal variation between a pair of TWTAs. There is in fact a seasonal variation that we have disposed of by looking at the worst day of the year. All the TWTAs will always have nearly the same temperature because of the heat pipes linking the north and south panels. Over the satellite's life, the TWTA temperature will rise due to the decreasing efficiency of the satellite's thermal subsystem, the effect of which is to gradually change the average daily P_{out}.

The CAMP's temperature-compensation circuit is intended to counteract the TWTA's temperature sensitivity for ALC mode. As discussed in the previous section on the CAMP, the compensation causes the TWTA's input drive to go up by 0.5 dB over the CAMP's operating temperature range of 65°C. For the few TWTAs whose P_{out} goes up with temperature, the CAMP's compensation will be in the wrong direction.

In our example, each TWTA's EOL P_{out} relative to BOL P_{out} can be computed from the formula given below:

$$\text{EOL } P_{\text{out}} \text{ rel to BOL } P_{\text{out}} = (35/85)[P_{\text{sat}\Delta} + gs(\text{OBO})(P_{\text{ss}\Delta} - P_{\text{sat}\Delta})] + (35/65)0.5\, gs(\text{OBO})$$

where gs(OBO) is TWTA gain slope in dB/dB at the operating output backoff (OBO) and $P_{\text{sat}\Delta}$, $P_{\text{ss}\Delta}$ are the TWTA's measured P_{out} variation at saturation and small signal, respectively, over 0–85°C. When the TWTA is saturated the function $gs(\text{OBO}) = 0$, and for small signal it equals 1. The formula can yield positive and negative values, depending on the parameter values. On the flight set of TWTAs, measured $P_{\text{sat}\Delta}$ ranged from -0.08 to -0.04 dB/dB, and measured $P_{\text{ss}\Delta}$ from -1.0 to -0.4 dB/dB. The 35° in the formula is the EOL temperature drift from BOL temperature, from Table 12.3. The 85° is the TWTA's operating temperature range. This formula does not go into the variation but affects the nominal value of P_{out}. If the lifetime requirement is "average over life" half of this difference should be used, while for "at EOL" the whole difference is used.

12.3.3.2 P_{out} Variation Due to Aging and Radiation

The TWTA's output power has an unknowable drift over life, due to aging and radiation that may be either positive or negative. Half the values of drift are beneficial to C and half are detrimental. The size of the drift at EOL is uniformly distributed between zero and the maximum positive EOL drift. We assume that the signals are not allowed to have their gain setting commanded during life; then the P_{out} variation over life corresponds to the small-signal gain drift over life.

Table 12.6 shows examples of the lifetime drift in P_{out} due to one TWTA's P_{out} drift from aging and radiation. For two signals, it is likely that either the drifts are in the same direction or are independent. Independence is worse so we assume that. The pdfs of one and of the sum of two independent drifts averaged over life are given in Appendix 8.A.7.

12.3.4 Measurement Uncertainty in P_{out}

We assume in our example that the size of P_{out} measurement error is to be taken into account (Section 8.3.2.4). The actual error is unknowable. Its probability distribution is approximately Gaussian with zero mean. Half the values of measurement error are beneficial to C and half are detrimental.

TABLE 12.6 C Variation due to TWTA's P_{out} Aging and Radiation Variation (Example)

	Lifetime Drift Unrelated to Temperature
	At EOL, in general form of a random number approximately uniformly distributed over $-L_2$ to L_2 where L_2 is
TWTA at saturation	0.05 dB
TWTA at small-signal	0.75 dB

TABLE 12.7 C Variation due to P_{out} Measurement Uncertainty (Example)

Constant but Unknown Offset
In general form of Gaussian-distributed number with mean 0 and std. dev. Q_2 where Q_2 is
0.125 dB

Table 12.7 shows an example of the uncertainty in a P_{out} measurement. The uncertainties of two measurements are assumed to be independent (worse than positively correlated), so the variation in the difference of two measurement uncertainties is the sum of the two individual, independent random numbers.

12.3.5 Variation Contribution from OMUX

The OMUX's insertion-loss change with temperature is a known function of temperature, only 0.1 dB over the operating temperature range of 80°C. The EOL diurnal swing has the panel-illumination pdf given in Appendix 8.A.6. The swing's limits are only ±0.03 dB, which is negligible. The situation between two OMUXes is shown in Table 12.8. If they are on different panels the delta's swing limits are ±0.07 dB, with the pdf given in Appendix 12.A.2.

At the point in time that we are assuming, manufacturing tolerance in the OMUX is also a contributor to uncertainty. An example of this uncertainty is given in Table 12.9. The uncertainties of two OMUXes are assumed to be independent (worse than positively correlated), so the uncertainty in the difference of two tolerances is the sum of the two individual, independent random numbers.

TABLE 12.8 C/I_{self} Variation due to OMUXes' Delta Diurnal Insertion Loss Variation (Example)

	Delta's Diurnal Swing
	In general form of sinewave with limits ±H where H is
OMUXes on same panel	0 dB
One OMUX on east panel, other on west	0.07 dB

TABLE 12.9 C Variation due to OMUX's Manufacturing Tolerance (Example)

Constant but Unknown Offset
In general form of Gaussian-distributed number with std. dev. Q_3 where Q_3 is
0.1 dB

TABLE 12.10 *C* Variation due to Manufacturing Tolerance of Other Post-TWTA Hardware Besides OMUX (Example)

Constant but Unknown Offset
In general form of Gaussian-distributed number with std dev Q_4 where Q_4 is
0.1 dB

12.3.6 Variation Contribution from Other Post-TWTA Hardware Besides OMUX

The post-TWTA hardware besides the OMUX consists of waveguide, possibly a few payload-integration components, and possibly another filter. At the point in time that we are assuming, manufacturing tolerance is the main contributor to uncertainty, and a small diurnal swing is comparatively negligible. An example is given in Table 12.10.

12.3.7 Summary of Repeater Units' Variation and Nominal-Value Contribution

For our example, Table 12.11 summarizes the constituents of the three components of the *C* variation. The three components as analyzed are constant but unknown offset,

TABLE 12.11 Summary of Constituents of Repeater-Caused *C* Variation (Example)

	Constituents of Repeater-Caused *C* Variation		
Unit	Constant but Unknown Offset	Diurnal Swing	Lifetime Drift Unrelated to Temperature
CAMP	Random number uniformly distributed over $-Q_1$ to $+Q_1$ (*settability*)	Negligible	At EOL, random number uniformly distributed over $-L_1$ to $+L_1$ *(aging and radiation)*
TWTA	N/A	0	At EOL, random number uniformly distributed over $-L_2$ to $+L_2$ (*aging and radiation*)
Power measurement	Random number Gaussian- distributed with std. dev. Q_2 (*msmt uncertainty*)	N/A	N/A
OMUX	Random number Gaussian-distributed with std. dev. Q_3 (*manuf tolerance*)	Negligible	N/A
Other post-TWTA hardware	Random number Gaussian-distributed with std. dev. Q_4 (*manuf tolerance*)	Negligible	N/A

diurnal variation, and lifetime drifts unrelated to temperature. The constant but unknown offset has four constituents (from CAMP settability, power-measurement uncertainty, and manufacturing tolerances in OMUX, and other post-TWA hardware) that are independent of each other. The diurnal variation is negligible. Lifetime drift has two constituents (from aging and radiation) that are uncorrelated with each other (the pdf of their sum averaged over life is given in Appendix 8.A.7).

For our example, each constituent of the C variation is independent of the corresponding constituent of the I_{self} variation except for the diurnal swing of OMUX insertion loss.

The only nonzero contributor to δ in the nominal value calculation (see Section 12.3.1) is the TWTA's P_{out} sensitivity to the rise in average TWTA temperature.

12.3.8 Repeater-Caused Variation and Nominal Value of C and C/I_{self}

We now form the composite repeater-caused variation of C and C/I_{self}. The three variation components as defined are mutually independent. We characterize the composite variation by its standard deviation. In what follows, we assume there is only one significant interferer, but to treat more the standard deviations for all would be rss'ed together.

We also characterize the nominal value about which the variation is centered in terms of δ.

12.3.8.1 Repeater-Caused C Variation and Nominal Value

The first component of the C variation due to the repeater, the unknown constant offset, is characterized as follows:

$$\begin{aligned}&\text{Constant but unknown component of } C \text{ variation in dB}\\ &\quad\approx \text{Random number with std. dev. } \sqrt{\tfrac{1}{3}Q_1^2 + \textstyle\sum_{i=2}^{4} Q_i^2}\end{aligned}$$

The second component, diurnal swing, is negligible. The third component, lifetime drift, is characterized as follows:

$$\begin{aligned}&\text{Lifetime drift component of } C \text{ variation in dB}\\ &\quad\approx \text{Random number with std. dev. } \sqrt{\tfrac{1}{9}\left(L_1^2 + L_2^2\right)}\end{aligned}$$

(If the lifetime condition were "at EOL" instead of "average over life," the divisor in the square root would be 3 instead of 9.) Therefore, the composite C variation due to the repeater is characterized as follows:

$$\begin{aligned}&\text{Variation in dB} \approx \text{Random number with std. dev.}\\ &\qquad \sigma_r = \sqrt{\tfrac{1}{3}Q_1^2 + \textstyle\sum_{i=2}^{4} Q_i^2 + \tfrac{1}{9}\left(L_1^2 + L_2^2\right)}\end{aligned}$$

The parameter δ equals half the drift over life from the TWTA's P_{out} sensitivity to the rise in average TWTA temperature.

TABLE 12.12 Parameter Values for Calculating Repeater-Caused C and I_{self} Variations (Example)

Variation Constituent Source	Parameter	Value for C (dB)	Value for I_{self} (dB)
CAMP settability	Q_1	0.25	0
CAMP aging and radiation	L_1	0.40	0.10
TWTA aging and radiation	L_2	0.75	0.05
Power-measurement uncertainty	Q_2	0.125	0.125
OMUX manufacturing tolerance	Q_3	0.10	0.10
Other post-TWTA hardware manufacturing tolerance	Q_4	0.10	0.10

Example: Suppose that the signal of interest has a well backed-off TWTA and an OMUX on the east panel, and that the strongest interferer has a saturated TWTA and an OMUX on the west panel. Then the parameter values for calculating the C and I_{self} variations are given in Table 12.12. Table 12.13 shows the terms in the sums whose square roots are the standard deviations. The standard deviations of the resultant variations are $\sigma_{r1} = 0.37$ and $\sigma_{r2} = 0.19$, respectively, where the subscript 1 means for C and 2 means for I_{self}. If the C variation is Gaussian then 95.4% of the time C will lie within $\pm 2\sigma_{r1}$ of its nominal value (Appendix 8.A.5), and similarly for I_{self}. Suppose that the parameter values for calculating the C and I_{self} EOL values relative to BOL are given in Table 12.14. Then the values of δ_1 and δ_2 are -0.07 and -0.02, respectively.

TABLE 12.13 Repeater-Caused C and I_{self} Variation Terms (Example)

Variation Constituent Source	Parameter	Value for C (dB)	Value for I_{self} (dB)
CAMP settability	$\frac{1}{3}Q_1^2$	0.021	0
CAMP aging and radiation	$\frac{1}{9}L_1^2$	0.018	0.001
TWTA aging and radiation	$\frac{1}{9}L_2^2$	0.063	0
Power-measurement uncertainty	Q_2^2	0.016	0.016
OMUX manufacturing tolerance	Q_3^2	0.010	0.010
Other post-TWTA hardware manufacturing tolerance	Q_4^2	0.010	0.010
Sum	σ_r^2	0.137	0.037

TABLE 12.14 Parameter Values for Calculating Repeater-Caused C and I_{self} EOL Values Relative to BOL (Example)

Parameter	Value for C (dB/dB)	Value for I_{self} (dB/dB)
$P_{sat\Delta}$	-0.06	-0.08
$P_{ss\Delta}$	-1.0	-0.7

12.3.8.2 Repeater-Caused C/I$_{self}$ Variation and Nominal Value The composite C/I_{self} variation due to the repeater is approximately given by

$$C/I_{\text{self}} \text{ variation in dB} \approx \text{Random number with std. dev. } \sqrt{\sigma_{\text{r1}}^2 + \sigma_{\text{r2}}^2 + \tfrac{1}{2}H^2}$$

where σ_{r1} and σ_{r2} equal σ_{r} for the signal of interest and the interferer, respectively, and $\pm H$ are the limits of the diurnal swing between the two OMUXes. When the OMUXes are on the same panel $H^2/2 = 0$, and when they are on different ones it is 0.002. Compared to $\sigma_{\text{r1}}^2 + \sigma_{\text{r2}}^2$ both values are negligible. Thus, the C, I_{self}, and C/I_{self} variations have no significant temperature-sensitive components, at least in our example. Composite C/I_{self} variation due to the repeater is then approximately given by

$$C/I_{\text{self}} \text{ variation in dB} \approx \text{Random number with std. dev. } \sqrt{\sigma_{\text{r1}}^2 + \sigma_{\text{r2}}^2}$$

The parameter δ in the nominal value calculation equals the difference in δs:

$$C/I_{\text{self}} \text{ average value at middle of life rel to BOL in dB} = \delta_1 - \delta_2$$

Example: Suppose that the signal of interest and the strongest interferer have the characteristics from the previous example. The C/I_{self} variation has standard deviation of 0.42 dB. If the variation is Gaussian then 97.7% of the time C/I_{self} will be within ± 0.84 dB of nominal. In regard to the nominal value, we have $\delta_1 - \delta_2 = -0.05$ dB.

12.4 COMBINING ANTENNA-CAUSED VARIATION INTO REPEATER-CAUSED VARIATION

In this section, we fold the payload antenna-caused variation into the repeater-caused variation of C and C/I_{self} and similarly for the nominal values. The complete list of payload sources of C and C/I_{self} variations is given in Table 12.15. They are approximately independent.

The antenna makes no contribution to δ in the nominal value calculation (Section 12.3.1) because the antenna's performance does not vary from year to

TABLE 12.15 Independent Components of C Variation due to Payload, as Analyzed

Repeater
Antenna-gain inaccuracy
Antenna-pointing error

year. So the composite payload δ is simply the value from the repeater, given in Section 12.3.8.1. The average performance of the payload over any year of life is based on antenna gain that has been averaged over antenna-pointing error.

12.4.1 Contribution from Antenna-Gain Inaccuracy

For the point in the build at which we are doing analysis, antenna-gain inaccuracy comes from manufacturing tolerance and computer-model error. On orbit, it will include thermal distortion.

The C variation contribution from antenna-gain inaccuracy has a Gaussian distribution:

$$C \text{ variation in dB from antenna gain inaccuracy} \approx \text{Gaussian random number with std. dev. } \sigma_g$$

On orbit the thermal distortion effect is different depending on whether the antenna is in the sun, partly in the sun, or in the dark, but in any case the term is not the largest contributor to overall antenna-gain loss (Schennum, 2011), so we simplify and use the worst-case value.

In regard to C/I_{self} variation, we need to think about how C and I_{self} are correlated. There seems to be no reason why they would be negatively correlated, so we make the worst-case assumption that they are independent.

12.4.2 Contribution from Antenna-Pointing Error

When the antenna does not autotrack, antenna-pointing error comes from spacecraft-attitude error and orbital motion, and to a lesser extent from initial antenna-boresight misalignment and from deformations from mechanical and thermal constraints (Maral and Bousquet, 2002). Orbital inclination and eccentricity vary with time. Nonzero inclination causes a north–south displacement with a 24 h period, while nonzero eccentricity causes an east–west displacement also with a 24 h period, for a GEO (Maral and Bousquet, 2002).

When the antenna does autotrack, antenna-pointing error does not depend on spacecraft-attitude or orbital motion, to first order.

Antenna pointing has two physically orthogonal dimensions whose errors are statistically dependent, both generally assumed to be Gaussian, with different nonzero means and different standard deviations (Maral and Bousquet, 2002).

Antenna gain in dB is a function of the two-dimensional pointing error. The gain is neither necessarily symmetric about zero pointing error nor does it necessarily have the same characteristics in the two dimensions. Antenna-gain variation is not generally Gaussian-distributed. The variation is relative to the gain obtained by averaging over antenna-pointing error.

The following is perhaps a good way to approximate the standard deviation of the C variation in dB, for a given location. Now, along the gain gradient at the location the gain changes the most, and in the orthogonal direction the gain does not

change. Using Gauss–Hermite integration (Section 8.5) over the Gaussian pointing error along the gradient, we calculate the rms gain variation in dB. Then we make the approximation that, over the two-dimensional pointing error, the standard deviation is the rms along the gradient, divided by the square root of two. Let us write

$$\begin{aligned} &C \text{ variation in dB from antenna pointing error} \\ &\quad \approx \text{Random number with std. dev. } \sigma_{\mathrm{p}} \approx \sqrt{\tfrac{1}{2}}\,\sigma_{\mathrm{p\ gradient}} \end{aligned}$$

For C/I_{self} there are two cases, namely when the two beams have uncorrelated pointing errors and when the two beams have basically the same pointing error at any given time. They have the same pointing error if they are on the same reflector. Also, they have basically the same pointing error if they are on different antennas that either do not track or change their pointing more slowly than the spacecraft attitude error changes.

The uncorrelated case is easy. The standard deviation is the rss of the two separate ones:

$$C/I_{\mathrm{self}} \text{ variation in dB} \approx \text{Random number with std. dev. } \sqrt{\sigma_{\mathrm{p1}}^2 + \sigma_{\mathrm{p2}}^2}$$

where the subscript 1 is for the signal of interest and 2 is for the interferer.

The case when the two beams off-point together is more complicated. We must keep in mind that if the interfering beam is a spot beam, at the location of interest the interferer is either on the steep side of the mainlobe of its antenna pattern or on one of the sidelobes (Section 11.5.3). Even if it is on a sidelobe, the pattern rolloff is probably sharper than that of the signal of interest, which is on the mainlobe of its antenna pattern.

Thus, if the interfering beam is a spot beam, the C/I_{self} gain gradient is approximately the gain gradient of the interferer. If the interfering beam is not a spot beam, the C/I_{self} gain gradient is approximately the gain gradient of the signal of interest. Now, it is possible that the other beam's gain gradient is orthogonal. (If it is in the same direction, it will be less steep and fairly negligible.) Therefore, a rough approximation and/or worst case is that, over the two-dimensional pointing error, the gain variation's standard deviation is the rms of the individual standard deviations along their gradients, divided by the square root of two. So we have the same approximation as above.

12.4.3 Payload-Caused *C* Variation

The composite variation in C from all of the payload is as follows:

$$C \text{ variation in dB} \approx \text{Random number with std. dev. } \sigma = \sqrt{\sigma_{\mathrm{r}}^2 + \sigma_{\mathrm{g}}^2 + \sigma_{\mathrm{p}}^2}$$

12.4.4 Payload-Caused C/I_{self} Variation

We have shown that with minor worst-case assumptions, the component of C variation due to the repeater is uncorrelated with the corresponding component of I_{self} variation, and similarly for the components due to the antenna. Therefore, the composite variation in C/I_{self} from all of the payload is as follows:

$$C/I_{\text{self}} \text{ variation in dB} \approx \text{Random number with std. dev. } \sqrt{\sigma_1^2 + \sigma_2^2}$$

where the subscript k is 1 for C and 2 for I_{self}.

Let us generalize this to the case where there is more than one significant interferer, so

$$I_{\text{self}} = \sum_{l=1,n} I_l$$

The Taylor series for a linear approximation to a function of multiple variables is as follows:

$$g(\eta, \mu_1, \ldots, \mu_n) \approx g(\eta_{\text{nom}}, \mu_{1\text{nom}}, \ldots, \mu_{n\,\text{nom}}) + (\eta - \eta_{\text{nom}}) \left.\frac{\partial g}{\partial \eta}\right|_{\text{nom}} + \sum_l (\mu_l - \mu_{l\,\text{nom}}) \left.\frac{\partial g}{\partial \mu_l}\right|_{\text{nom}}$$

where

$$\eta = 10 \log C$$

$$\mu_l = 10 \log I_l$$

$$g(\eta, \mu_1, \ldots, \mu_n) = 10 \log (C/I_{\text{self}})$$

where $\log \triangleq \log_{10}$.

Then

$$\begin{aligned} C/I_{\text{self}} \text{ variation in dB} &\triangleq 10 \log \left(C / \sum_l I_l \right) - 10 \log \left(C / \sum_l I_l \right)_{\text{nom}} \\ &\approx (10 \log C - 10 \log C_{\text{nom}}) - \sum_l \alpha_l (10 \log I_l - 10 \log I_{l\text{nom}}) \\ &= (C \text{ variation in dB}) - \sum_l \alpha_l (I_l \text{ variation in dB}) \end{aligned}$$

where $\alpha_l = I_{l\text{nom}}/I_{\text{self nom}}$. The composite variation in I_{self} can be characterized as follows:

$$I_{\text{self}} \text{ variation in dB} \approx \text{Random number with std. dev. } \sigma_2 = \sqrt{\sum_l \alpha_l^2 \sigma_{l2}^2}$$

Thus, the composite variation of C/I_{self} in dB can be written in the same form as for one interferer.

12.5 PAYLOAD-CAUSED VARIATION OF $C/(I+N)$

Now we extend the analysis of the previous section on C and C/I_{self} to $C/(I_{\text{self}}+N)$, where N includes I_{other} (Sections 12.1 and 12.2). N and I_{other} are given and are referred to the payload-antenna output, as C and I_{self} are. N is calculated from space loss and the ground receiver's G/T.

Making a calculation similar to the one in the previous section, we find that the composite variation of $C/(I_{\text{self}}+N)$ in dB can be characterized in the following way, where $\beta = I_{\text{self nom}}/(I_{\text{self nom}}+N)$:

$$C/(I_{\text{self}}+N)\text{variation in dB} \approx \text{Random number with std. dev.}\sqrt{\sigma_1^2+\beta^2\sigma_2^2}$$

where subscript 1 means for C and 2 means for I_{self}. The composite nominal value of $C/(I_{\text{self}}+N)$ in dB is the following:

$$[C/(I_{\text{self}}+N)]_{\text{nom}} \text{ in dB } \approx (C_{\text{nom}} \text{ in dBW}) - \beta(I_{\text{self nom}} \text{ in dBW})$$

Example: We consider only the repeater-caused component of variations and use the numbers from Section 12.2.3. We further assume that $N = I_{\text{self nom}}$, which means $\beta = 0.5$. Then the standard deviation is 0.38 dB.

12.6 COMBINING ATMOSPHERE-CAUSED VARIATION INTO PAYLOAD-CAUSED VARIATION

This section combines a statistical model of atmospheric attenuation (Section 11.3.4.2) with the probabilistic model of payload performance. The usual EIRP requirements are assumed to have been replaced by a long-term link-availability requirement or equivalently by a maximum long-term link-outage probability q_{spec} in percentage.

Two aspects of the atmosphere are most likely to cause significant link attenuation at frequencies above 10 GHz: the dominant one is rain and the other is atmospheric gas, especially water vapor.

Link margin is the ratio of atmosphere-attenuated $C/(I_{\text{self}}+I_{\text{other}}+N)$ to the threshold value r_{th} required by the receiver for good performance. As in the previous section, we fold I_{other} into N. The atmosphere attenuates C and I_{self} by the same amount $a>1$, but N (including the part from I_{other}) is not affected. So the link margin is as follows:

$$m_{\text{link}} = m_{\text{link}}(C,\, I_{\text{self}},\, a) = \frac{C/a}{I_{\text{self}}/a + N} \times \frac{1}{r_{\text{th}}}$$

The link margin is greater than or equal to 1 (non-negative, if in dB) if and only if the link is available. The link margin is less than 1 (negative, if in dB) if and only if the link is suffering an outage. When the link margin is greater than 1 (positive, if in dB), there is some room for C, I_{self}, and a to vary.

The **atmospheric margin** m (not in dB) is something different from link margin. It is the atmospheric attenuation at which the link margin equals 1 (zero, if in dB):

$$m = m(C, I_{\text{self}}) = \frac{r_{\text{th}}^{-1} - (C/I_{\text{self}})^{-1}}{(C/N)^{-1}}$$

It depends on the instantaneous values of C and I_{self}. Define $M = 10 \log m$, the atmospheric margin in dB.

The long-term probability of link outage is the link outage averaged over the probability distributions of C, I_{self}, and atmospheric attenuation a. We need the nominal values and standard deviations of C and I_{self}, all in dBW. The computational technique is that we set C and I_{self} to values in their ranges, then conditioned on these values we compute the link-outage probability, then we average over the C and I_{self} probability distributions. To make the problem tractable, we assume that the payload's performance variation in dB is Gaussian-distributed. It is unknowable how good this approximation is because there are so many contributors to the variation, each with not-so-well-known probability distributions, but it is most likely a fairly good approximation. Partial precedence for this approach is the computation of antenna-pointing error (see Section 11.3.4). We use Gauss–Hermite integration (Section 8.5) to perform the two-dimensional averaging. The steps for computing the link-outage probability at the location are as follows:

- Decide how many points to use for each one-dimensional Gauss–Hermite integration (try 4 and 6). Obtain the corresponding set of factors η_i and the set of weights w_i.
- Compute C_{nom}. Compute σ_1 characterizing the payload variation (Section 12.4.3).
- Compute the discrimination factor (Section 11.5) that C has against each of its possible significant interferers. Identify the significant ones and compute the α_l according to Section 12.4.4. Compute I_{nom} and σ_2.
- Compute and store the constants for evaluating the inverse atmospheric-attenuation function (Appendix 12.A.1.2).
- Compute all $C_i = C_{\text{nom}}10\text{^}(\eta_i\sigma_1/10)$ and all $J_i = I_{\text{nom}}10\text{^}(\eta_i\sigma_2/10)$. For each index pair (i,l), compute the atmospheric margin $M_{il} = M(C_i,J_l)$. For atmospheric attenuation $A = M_{il}$, evaluate the inverse atmospheric-attenuation function to yield the link-outage probability q_{il} (see Appendix 12.A.1.2). Perform the two-dimensional weighted averaging to obtain the long-term average link-outage probability q_0:

$$q_0 = \sum_i w_i \sum_l w_l q_{il}$$

12.7 OPTIMIZING MULTIBEAM-DOWNLINK PAYLOAD SPECIFIED ON LINK AVAILABILITY

We wish to optimize a multibeam-downlink payload that has been specified by a minimum long-term link availability. A large set of locations is given on which the specification must hold (see Section 12.2). There are so many different constraints on the payload design that the optimization must be done iteratively. We sketch out how to iteratively determine a set of C_{nom}, one for each location, which meets the availability requirement while the signal levels vary in the manner discussed in the preceding sections. Once the final set of C_{nom} is found, the corresponding set of C_{BOL} is calculated from the difference between C_{nom} and C_{BOL} at every location.

The payload does not change configuration between iterations but only at the beginning of a set of iterations. "Configuration" means the CAMPs, HPAs, and antennas. The idea is that a provisional payload configuration is set, a few iterations are done here until a problem is seen with the configuration, the payload is reconfigured, and so on. "I_{self}" is written simply "I" here.

a. To do once:
   i. Decide how many points to use for the Gauss–Hermite integrations. Obtain the corresponding set of factors η_i and the set of weights w_i.
   ii. Set the iteration adjustment factor ξ_0 for upward adjustments of C_{nom}, referred to in the iteration details presented in Appendix 12.A.1.3. It is a positive number up to 1. A value of 0.5 may be good.

b. Steps to do once for every location, for the given configuration:
   i. Obtain an initial C_{nom}, for example, the one that exactly meets link availability without consideration of self-interference. Compute and store σ_1 (Section 12.4.3).
   ii. Identify and store the small number of indices of the potentially significant interfering beams.
   iii. Compute and store the constants for evaluating the approximate forward and the inverse atmospheric-attenuation functions (Appendix 12.A.1.2).
   iv. Initialize the "previous" iteration-quality measure to zero.

c. Steps to do at every iteration of finding a whole new set of C_{nom}:
   i. For every location,
      1. Check its small set of potentially significant interfering beams. Compute the discrimination factor that C has against each of them. Identify the significant ones and compute the α_I. Compute I_{nom} and σ_2.
      2. Compute and store $C_{\text{nom new}}$ (Appendix 12.A.1.3).
      3. If $C_{\text{nom new}}$ causes the HPA to exceed its capacity, reconfigure the payload and go to step **b**.
      4. Otherwise, replace C_{nom} with $C_{\text{nom new}}$.
   ii. Compute the iteration-quality measure of this iteration, which is the number of locations that meet requirements. Check to see that it is better

than the value from the previous iteration. If not, adjust the factor ξ_0 and start this iteration again. Set the "previous" value of the measure to the current value.

iii. Make a plot of all the C_{nom}.

d. After a few iterations, if C_{nom} has been consistently rising for all locations in some subset then the discrimination of each against the others is not sufficient. Increase the discriminations in at least one of the discrimination dimensions (frequency band, antenna pattern, polarization) and go to step **b**. Also check all the C_{nom} for possibly exceeding the maximum-allowed flux density on the earth.

e. After the iterations have converged, steps to do for every location:

i. Compute $P_{\text{sat}\Delta}$ and $P_{\text{ss}\Delta}$ and from them compute C_{nom} relative to the C_{BOL} without pointing error. Compute the increase required to C to compensate the average loss from antenna-pointing error. Compute C_{BOL} from C_{nom} and both adjustments.

If convergence seems impossible, consider reducing the C variations by tightening up unit specifications.

APPENDIX 12.A

12.A.1 Iteration Details for Optimizing Multibeam Payload Specified on Link Availability

12.A.1.1 Approximate Rain-Attenuation Function and Its Inverse The rain attenuation R for a location is a function of the exceedance probability p (Section 11.3.6), that is, $R = R(p)$. R is in dB and p is in %. Since both $R(p)$ and its inverse function $p(R)$ will be evaluated repeatedly, we wish to replace them with efficient approximations.

For the ITU rain-attenuation model and therefore presumably for other models, $\log R$ and $\log p$ are nearly linearly related over a not-so-small range. The range of interest is for p near the specified long-term link-outage probability q_{spec}. With the actual rain-attenuation model, we compute and store $R_{\text{spec}} = R(q_{\text{spec}})$. Then for p near q_{spec} we have the following good approximation:

$$\log R \approx B'' + D' \log p \quad \text{where } D' = \left[\frac{d \log R(p)}{d \log p}\right]_{p=q_{\text{spec}}}$$

$$B'' = \log B' \qquad \text{where } B' \triangleq R_{\text{spec}} q_{\text{spec}}^{-D'}$$

The forward relationship is the following:

$$R = R(p) \approx B' p^{D'}$$

The inverse relationship is the following:

$$p = p(R) \approx BR^D \quad \text{where } D \triangleq 1/D' \text{ and } B \triangleq R_{\text{spec}}^{-D} q_{\text{spec}}$$

We compute and store D' and B' for use in the forward-function approximation D and B for use in the inverse function.

If the rain model changes expression at $p = q_{\text{spec}}$ then the tangent line may not be defined, that is, be different on the lower side and the upper side of q_{spec}. Then we could define two tangent lines and know that they apply in different regions. Or it may work to just define one tangent line with the slope equal to the average of the slopes on either side.

12.A.1.2 Atmospheric Attenuation Function and Its Inverse The atmospheric attenuation A for a location is a function of the long-term link-outage probability q (Section 11.3.6). We assume that the other atmospheric attenuations besides the rain attenuation R are constant. We obtain and store them. Then A and R are related one-to-one, that is, $A = A(R)$ and $R = R'(A)$, respectively. The forward atmospheric-attenuation function is $A = A(R(q))$ and the inverse function is $q = p(R'(A))$, where $R(p)$ and $p(R)$ are given in the preceding section.

12.A.1.3 Details of Iteration Compute all $C_i = C_{\text{nom}}10\text{^}(\eta_i\sigma_1/10)$ and all $J_i = I_{\text{nom}}10\text{^}(\eta_i\sigma_2/10)$.

For each index pair (i,l), compute the atmospheric margin $M_{il} = M(C_i,J_l)$. For atmospheric attenuation $A = M_{il}$, evaluate the inverse atmospheric-attenuation function to yield the link-outage probability q_{il} (see Appendix 12.A.1.2). Perform the two-dimensional weighted averaging to obtain the long-term average link-outage probability q_0 for the current C_{nom}:

$$q_0 = \sum_i w_i \sum_l w_l q_{il}$$

Use the forward atmospheric-attenuation function (also Appendix 12.A.1.2) to compute the total atmospheric attenuation a_0 (not in dB) corresponding to q_0.

Now, we are trying to achieve a one-point-equivalent atmospheric margin equal to the attenuation at the specified outage probability, a_{spec}, but in this iteration we reached a_0. All we do during an iteration is calculate the new C_{nom} for the next iteration, so the adjustment we need, we can only attribute to the new C_{nom}. C_{nom} needs to be multiplied by approximately the factor a_{spec}/a_0. If we have too much atmospheric margin, we can decrease C_{nom} by the full amount without making a negative impact on other locations. On the other hand, if we do not have enough atmospheric margin then we have to increase C_{nom}, but since this may have a negative impact on other locations, we only increase it part way. In symbols, if $a_0 > a_{\text{spec}}$, set the iteration adjustment factor ξ to 1; otherwise set it to ξ_0. Then

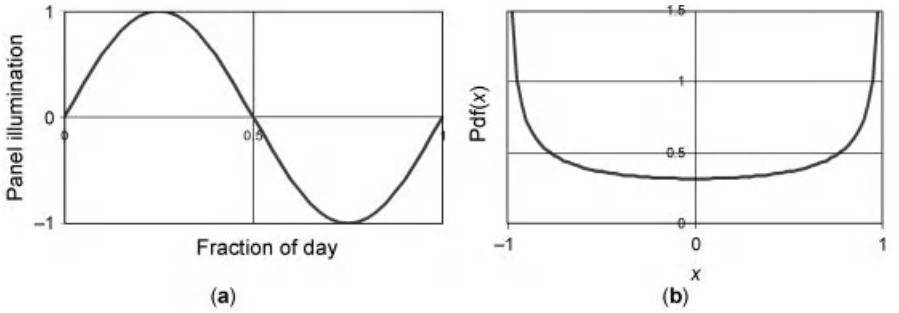


FIGURE 12.2 Difference in diurnal variation in east and west panel illumination: (a) difference in illumination over the day and (b) pdf.

compute and return $C_{\text{nom new}}$ for the next iteration:

$$C_{\text{nom new}} = C_{\text{nom}}\left(a_{\text{spec}}/a_0\right)^{\xi}$$

12.A.2 Pdf of Diurnal Variation in Delta of East and West Panel Illumination

In Appendix 8.A.6, we saw how the amount of illumination that the east or west panel gets varies over the day. The panel is in the sun half the day and in the shade the other half. The amount of illumination starts at zero, acts like a sine function until it hits zero again, and then stays at zero for the second half of the day. Here we are interested in the delta of the diurnal illuminations of the east and west panels. One panel is in the sun while the other is in the shade. The delta diurnal illumination is plotted in Figure 12.2a, where the starting phase is arbitrary. The pdf for the delta variation is drawn in Figure 12.2b and stated as follows:

Pdf for diurnal variation in delta of east and west panel illumination, at x from -1 to 1, is $\left(\pi\sqrt{1-x^2}\right)^{-1}$, which has mean 0 and std. dev. $\sigma \approx 0.707$

The use of this pdf is to approximate the diurnal delta temperature of OMUXes and the resultant variation of delta insertion loss. The pdf's independent axis values would range from diurnal minimum delta to maximum.

REFERENCES

Maral G and Bousquet M (2002). *Satellite Communications Systems*, 4th ed. Chichester, England: John Wiley & Sons Ltd.

Schennum G, Space Systems/Loral (2011). Personal communication; Mar. 23.

CHAPTER 13

END-TO-END COMMUNICATIONS SYSTEM MODEL WITH FOCUS ON PAYLOAD

13.1 INTRODUCTION

This chapter presents ways to model the end-to-end communications system with focus on analyzing the payload's effect on performance. System performance is assumed to be evaluated in terms of symbol error rate (SER) or bit error rate (BER) (Section 10.5.4). Modeling may be done to help develop the payload specifications, to assess the impact of a payload parameter not meeting its requirement, to figure out what could be causing an on-orbit anomaly, or to try out a system upgrade on a test bed before committing to implementing it.

There are three general ways to model: by hand mathematically or with Microsoft's Excel, by software **simulation** on the computer, and by hardware **emulation**. It is possible and even sometimes advisable to combine techniques. For example, hand calculations may show the levels of interfering signals compared to the level of the signal of interest or how interfering signals can be modeled in a simplified way; hardware can be characterized and the measurements input to a simulation; or performance of the system in the nominal operating condition may be found by emulation, and this nominal performance used to calibrate a simulation, which is then used to assess system performance sensitivity to a parameter variation.

The rest of this chapter has three sections:

- *Section 13.2*: Considerations common to both simulation and emulation. Mathematical techniques for model simplification

Satellite Communications Payload and System, First Edition. Teresa M. Braun.

- *Section 13.3*: Considerations particular to simulation. Further mathematical techniques for simulation simplification
- *Section 13.4*: Considerations particular to emulation.

13.2 CONSIDERATIONS FOR BOTH SOFTWARE SIMULATION AND HARDWARE EMULATION

13.2.1 System Model

A model of the end-to-end satellite communications system with a bent-pipe payload consists of three main models, of the ground transmitter, the satellite payload, and the ground receiver. For a regenerative payload there would be four, as the payload would be split into a receiving and demodulating model and a modulating and transmitting model. The regenerative payload leads to a simpler end-to-end model, since there is no noise on the signal in the payload's high-power amplifier (HPA), so for the rest of this section we consider only the bent-pipe payload. We do not treat framing or software protocol issues.

Figure 13.1 gives block diagrams of typical models of the ground transmitter, payload, and ground receiver. There is both uplink and downlink noise, coming from the payload and ground receiver, respectively. The main nonlinear element in the system is the payload HPA in this typical case. The individual elements in the payload have been discussed in earlier chapters, and the elements on the ground in Chapter 10. In an emulation, all of the elements in the three models would be

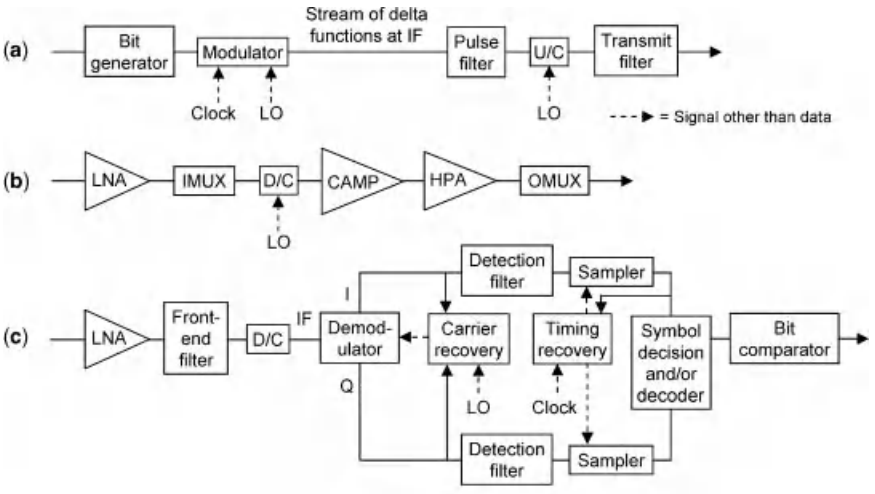


FIGURE 13.1 Typical system model with bent-pipe payload: (a) ground transmitter, (b) payload, and (c) ground receiver.

present. In a simulation they would be, too, but the representations would be more abstract (see Section 13.3.2).

Additional system elements sometimes must be modeled:

- Error-control coder and decoder
- Interfering signals entering low-noise amplifiers (LNAs)
- Interfering signals and spurious signals entering elsewhere in system
- Payload-unit performance variation with temperature and/or life
- Antenna noise
- Antenna pointing error
- Atmospherics on link(s)
- Doppler and Doppler rate on link(s)
- Nonlinear HPA in ground transmitter
- Limiter in payload front end
- Limiter in ground-receiver front end.

13.2.2 Know the Whole Communications System

It is crucial, even when the analysis focus is on the payload, to model every element of the end-to-end communications system accurately. If not, the system performance results can be so misleading as to show nothing about the actual system but something about a quite different system, and a false conclusion is drawn. The reason this can happen is that one aspect of the system can magnify the performance effect of another aspect or diminish it (Braun and McKenzie, 1984).

Assuming that the payload characteristics are known in detail, the rest of what has to be known (aside from the possible additional system elements listed above) includes the following:

- Uplink and downlink levels of signal of interest and the ground-receiver noise figure
- Modulation and demodulation schemes and detection filter. Large ground stations normally have excellent performance so can often be modeled as ideal, but this must be ascertained. Consumer equipment is typically not near-ideal. Some ground receivers may have "legacy" equipment, which is older and possibly optimized for an earlier transmitter or not up to current performance standards.
- Phase noise characteristics of ground LOs and clock jitter characteristics of ground clocks
- Carrier tracking and symbol tracking schemes
- All other ground filters
- Interference scenario
- What the coding, decoding, and interleaving schemes are, if any
- What the ground receiver's antenna noise is

- What the ground transmitter's and ground receiver's antenna-gain loss is from pointing error, if significant
- Locations of ground equipment, if atmospherics is important
- Nonlinear behavior of ground transmitter's HPA and ground receiver's front end, if significant.

All system elements are to be in nominal operation unless specifically a worst case is being studied. Thus, element specification data is not sufficient; measured data is required. The reason is again the magnifying or diminishing effect of one parameter on another.

13.2.3 What Results Modeling Can Provide

Neither simulation nor emulation, and certainly not hand analysis, can provide performance to match that of the actual analog hardware system being modeled. Simulation can never match hardware because hardware has second- and third-order behavior, including interactions, that is untested, unknown, and thus unmodeled. (The author remembers that in about 1994 we had gotten our first results on a new-technology testbed and the results did not match the customer's simulation. The customer was shocked and thought the hardware results were wrong and wanted us to work through Thanksgiving. This was the beginning of years of discussion about accurate simulation.) Emulation—using one set of hardware and test equipment and operating under one set of environmental conditions and having turned on the equipment so much time beforehand—can never match another set of hardware and test equipment, and so on, even if all is nominally identical. Measurement error alone will often obscure the "real results" by tenths of a dB (Chie, 2011).

So if simulation and emulation cannot match performance of the modeled system, what can they do? They are at their most useful in predicting sensitivity, namely, what the change in performance is when one system parameter is varied (basically a partial derivative). This is why the model must match the real system as closely as possible. Simulation and emulation can also possibly say that one version of unit x is better than another version of unit x.

Once the model is as realistic as possible, one still has to calibrate the model to get it operating with the same performance as the real system, for example, the same BER. Frequently the only way to model the otherwise unmodeled aspects is to inject more noise in an appropriate place.

13.2.4 Generating Symbol Stream Plus Noise

Generating the bit stream by a **maximal-length pseudo-noise (PN) sequence** (Jeruchim et al., 2000) is more efficient than by a random-number generator. A PN sequence is a periodic bit sequence whose autocorrelation function looks much like that of a randomly generated bit sequence. It is generated with a feedback shift register. Suppose the feedback shift register has m memory locations. After each shift, it computes a linear, base-two combination of some of the m bits in memory and

outputs that bit. The PN sequence generated has maximal length if all possible runs of 2^m bits are represented in the repeated sequence. However, the run of m zeros cannot be present, since then the whole sequence would be only zeros. So the PN sequence must be altered to add one more zero to the run of $m-1$ zeros.

The symbol stream is then generated from the bit stream, each symbol from M consecutive bits where M is 1 (for BPSK), 2, 3, or 4 (for 16QAM) in the modulation-symbol schemes presented in Section 10.3.3.1.

We need to determine how big m must be. Now, the satellite communications system has memory. Without regard to the HPA, phase error, timing jitter, spurious signals, and so on, the system from pulse filter through demodulator can be represented by one filter. Say its memory is n symbols long, which corresponds to Mn bits. All bit sequences of length Mn must be present in the modified PN sequence, so we need m to be at least a few times Mn. One way to find out what the symbol-memory length n is, is to send an impulse down the system without noise and look at the response in the time domain. The number n should be set to the length of the main content of the response. This method would work for a channel that is not highly nonlinear.

If the channel is highly nonlinear, a trial-and-error approach can be used, in which some value of m is tried and the SER is obtained, a larger value of m is tried and that SER is obtained, and if the answers are nearly the same the first m was big enough.

If the modeling is either by emulation or Monte Carlo simulation (Section 13.3.4) then the modified PN sequence is used over and over to generate the data. In the former case the hardware itself adds noise, while in the latter case "random" noise must be generated and added in. We investigate now how long the total noisy symbol sequence has to be in order for the SER obtained by emulation or Monte Carlo simulation to be a good approximation. A precondition for the following rule is that the symbol errors seen in the simulation must be independent of each other (Jeruchim et al., 2000). Suppose it is desired that the actual value of SER lie within a factor of $1/\rho$ to ρ about the simulated SER with 95% probability, where $\rho > 1$. Then 10 symbol errors are enough for ρ of 2 and 100 for ρ of 5/4 (Jeruchim et al., 2000).

However, because of the system memory, successive symbols are in fact correlated, so their errors are correlated. Now, for an uncoded link carrying data, not voice, the target SER may be about 10^{-5}, while for a coded link it may be 10^{-2} before decoding and a doubly coded link 10^{-1} before any decoding. (These SERs are called the "uncoded SER.") For SER of 10^{-5} and 10^{-2}, anyway, the errors are uncommon enough that the independent-error assumption is not far from correct. For the doubly coded case, it may be necessary to specifically model the outer (defined below in Section 13.3.7) encoder and decoder.

13.2.5 Modeling Other Signals Present

Other signals may come into the end-to-end system and degrade recovery of the signal of interest. These other signals may be self-interference, interference from other communications systems, other signals going through the HPA with the signal of interest, spurious signals from a frequency converter, and so on. For simplicity just in the rest of this chapter we call them all "interference."

Approximating interferers as additive noise, when appropriate, simplifies a system model significantly and cuts down on the computer run-time, without loss in accuracy. A good system modeler knows when it is appropriate, and this is what we will discuss right below.

One important thing to keep in mind when deciding how to model an interferer is what filtering the interferer will see compared to the signal of interest. The main filter is the detection filter, but other filters can be significant such as the output multiplexer (OMUX) or a post-HPA notch filter. If the interferer magnitude will be significantly reduced compared to the signal of interest or a strong interferer, the need for accuracy in the (first) interferer's model is reduced.

13.2.5.1 Applying Central Limit Theorem The Central Limit Theorem is one of the most useful theorems in communications analysis. It says that the sum of a number n of random variables (rvs) that are mutually independent and identically distributed (iid) becomes closer and closer to Gaussian as n becomes larger (Papoulis, 1984). Because the rvs are iid, the mean and standard deviation of the sum are respectively n times the mean and n times the standard deviation of one such rv.

The convergence to Gaussian is remarkably fast in most cases we come across, luckily for our analysis and simulation. Most of the rvs in our sums have either a Gaussian or a uniform pdf, and usually there is a mixture. We already know that the sum of Gaussian rvs is Gaussian, so let us look at the sum of rvs with uniform pdfs. We want to find out how many independent rvs have to be in the sum for the sum's pdf to be well approximated by Gaussian within $\pm 2\sigma$ of the mean value. Figure 13.2

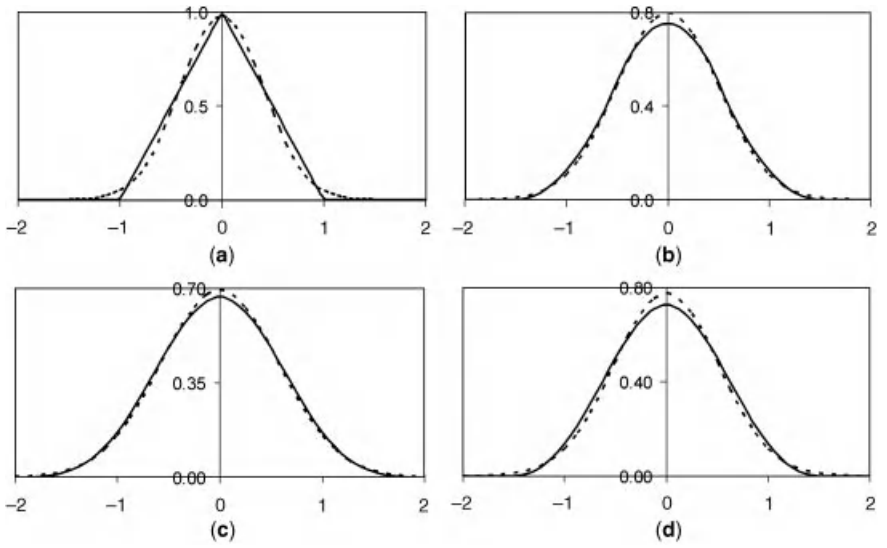


FIGURE 13.2 Pdfs of sums of uniformly distributed random variables: (a) sum of 2, (b) sum of 3, (c) sum of 4, and (d) sum of 3 of different pdfs. Gaussian pdfs shown with dashed lines.

TABLE 13.1 Integrals on $\pm 1\sigma$ and $\pm 2\sigma$ of Pdfs of Sums of Independent, Uniformly Distributed Random Variables

Random Variables in Sum	Integral on $\pm 1\sigma$	Integral on $\pm 2\sigma$
Two uniform, both on ± 0.5	0.65	0.966
Three uniform, all on ± 0.5	0.66	0.957
Four uniform, all on ± 0.5	0.67	0.957
One uniform on ± 0.5, another uniform on ± 0.35, and another uniform on ± 0.65	0.66	0.961
Gaussian	0.68	0.954

shows pdf plots for the following cases involving uniformly distributed rvs: (a) sum of two rvs each distributed on ± 0.5, (b) sum of three such, (c) sum of four such, and (d) sum of one distributed on ± 0.5, one on ± 0.35, and one on ± 0.65. Zero mean is assumed for all rvs. The standard deviation of the cases is 0.41, 0.5, 0.58, and 0.51, respectively. In each plot the Gaussian pdf of the same standard deviation is shown as a dashed line. The convergence to Gaussian with increasing n is visible in the progression of the plots (a), (b), and (c).

Table 13.1 shows the integrals of the same four sum pdfs over $\pm 1\sigma$ and $\pm 2\sigma$ as well as the integrals for Gaussian. Even for only three uniformly distributed rvs in the sum, the integrals are very close to the integrals for Gaussian. The case of three uniformly distributed rvs with different standard deviations, corresponding to plot (d) in Figure 13.2, is given in the table as well to show that even if the standard deviations of three rvs of the kind likely to be found in a payload analysis are somewhat different, the sum of the rvs is close to Gaussian. However, if one or two of the three are much larger than the others, Gaussian may not be a good approximation. This is something to watch out for.

An extreme but thereby revealing example is the sum of a few randomly phased cosines. The pdf of one cosine, shown in Figure 13.3a, could hardly be less Gaussian-like. However, the power of summing a few rvs to yield a Gaussian-like pdf is illustrated in Figure 13.3b, c, and d, for three, four, and five, respectively, in a sum. Note the different scales on the various plots.

The conclusion is that in most situations to be encountered in communications analyses involving the payload, the sum of as few as three iid random variables is approximated well on $\pm 2\sigma$ by a Gaussian random variable. Furthermore, when we do not know the individual pdfs accurately, which is often the case, there is even more reason to accept the approximation.

13.2.5.2 Approximating Sinewave Interferer by Noise-Like Interferer

We would like to know when a continuous-wave (CW) interferer can be well approximated by a noise-like interferer, since it is easier to model the latter. We assume the signal is QPSK with any pulse (Section 10.3.4) and the system is otherwise ideal except for additive white Gaussian noise and the interferer. The CW interferer is assumed to be

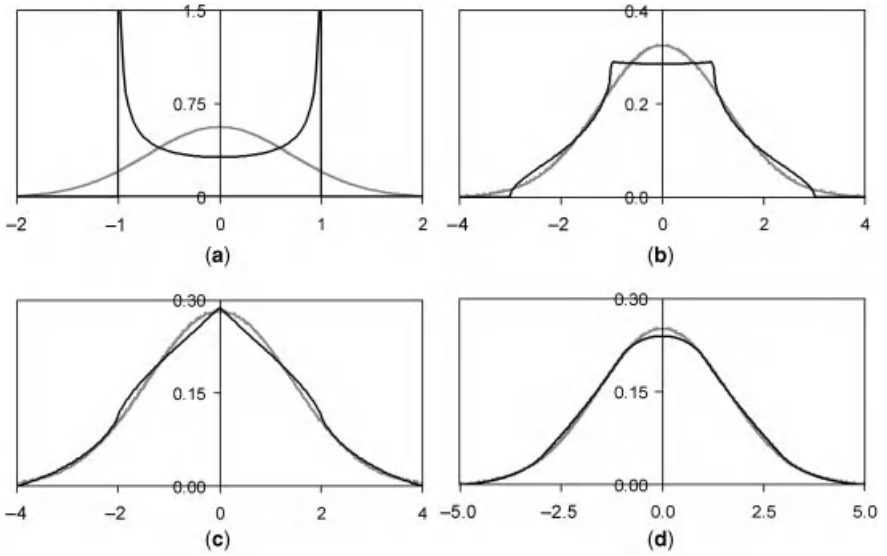


FIGURE 13.3 Pdfs of sums of independent, randomly phased cosines: (a) one, (b) sum of 3, (c) sum of 4, and (d) sum of 5. Gaussian pdfs of same standard deviations shown with gray lines.

very close to but not at the carrier, which is the worst case of a noncoherent CW interferer.

Figure 13.4 compares the BER degradation caused by the two interferers when they have the same *C*/*I*. The noise bandwidth for the noise interferer is taken as the modulation-symbol rate. For all the values of E_b/N_0 and *C*/*I* treated, the CW interferer is more benign. The spread between the degraded BER curves at a given *C*/*I* increases as E_b/N_0 increases. So for this particular CW interferer, modeling it as a little extra noise is a conservative approximation. For a different modulation, a similar analysis would have to be carried out.

Let us now look at a CW interferer that lies at the 3-dB-down point of the detection filter, and compare it to the noise-like interferer of the previous example. Degraded BER curves for this are shown in Figure 13.5. The CW interferer is considerably more benign than the noise interferer, at low *C*/*I* values. Similarly, a CW interferer located at the skirts of the detection filter will have almost no degrading effect. Sometimes we do not know at what frequency a CW interferer is, in which case we must think about where to place it in a model.

The noise approximation may be sufficient for CW interferers that come into the signal of interest after the payload's HPA or for weak interferers, but large CW interferers that come in before the HPA must be explicitly modeled through the HPA, as they create large IMPs. In fact, they can make carrier acquisition impossible or degrade carrier tracking.

13.2.5.3 Approximating Modulated-Signal Interferer by Noise-Like Interferer

We assume that neither the signal of interest nor the interferer has

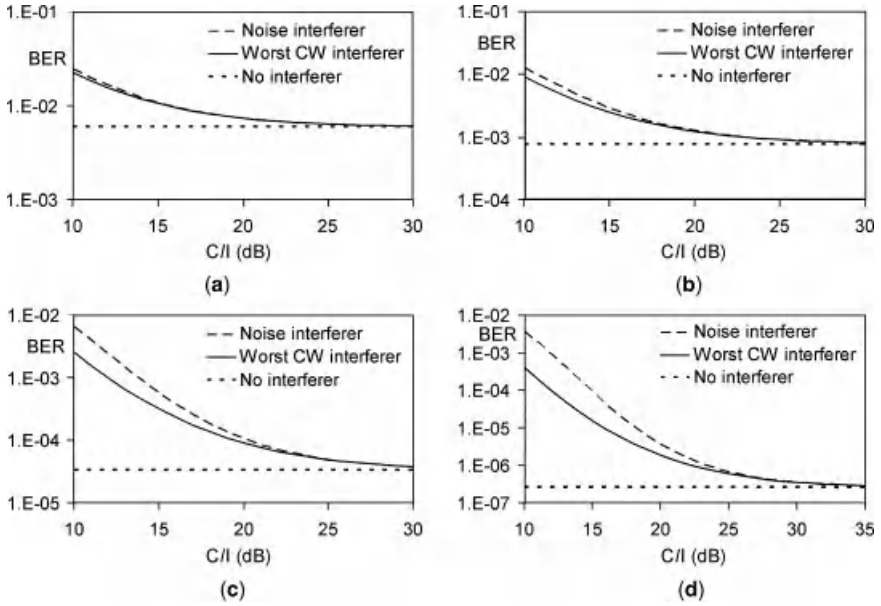


FIGURE 13.4 Comparison of worst noncoherent CW interferer and noise-like interferer effects on QPSK BER. (a) $E_b/N_0 = 5$ dB, (b) $E_b/N_0 = 7$ dB, (c) $E_b/N_0 = 9$ dB, and (d) $E_b/N_0 = 11$ dB.

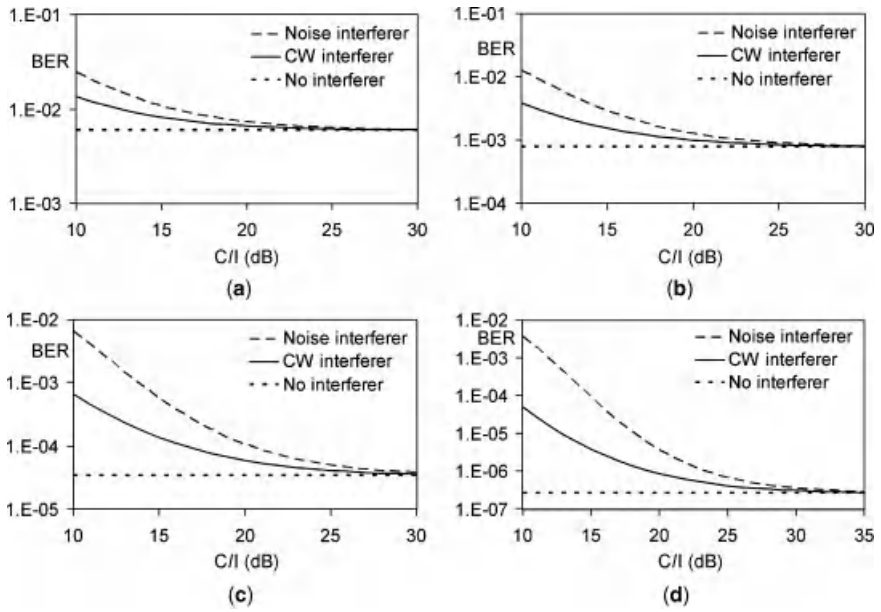


FIGURE 13.5 Comparison of half-power worst noncoherent CW interferer and noise-like interferer effects on QPSK BER. (a) $E_b/N_0 = 5$ dB, (b) $E_b/N_0 = 7$ dB, (c) $E_b/N_0 = 9$ dB, and (d) $E_b/N_0 = 11$ dB.

BPSK modulation. Then whatever portion of the interferer's signal spectrum is within the noise bandwidth of the detection filter of the signal of interest, averaged over that noise bandwidth, can be treated as an "I_0" level to be added to the N_0 level.

This approximation is sufficient for modulated-signal interferers that come into the signal of interest after the payload's HPA or for weak interferers, but large modulated-signal interferers that come in before the HPA must be explicitly modeled through the HPA.

13.3 ADDITIONAL CONSIDERATIONS FOR SIMULATION

Among the reasons for performing a simulation with focus on the payload are the following: (1) to help determine the payload specifications; (2) to assess the BER sensitivity to a range of values for a payload parameter to support a request for specification relief; and (3) to assess the BER impact of a payload parameter already known to be out of spec.

Simulation is often done with a software package such as the MathWorks' product Matlab. For simple simulation, PTC's product Mathcad is sufficient. Both of these packages operate alternately in the time and frequency domains as is convenient. Two new tools explained in the open literature may be useful. One is a demonstration tool developed by the Aerospace Corp., a hardware-accelerated simulation tool (HAST) consisting of field-programmable gate arrays (FPGAs) to perform computation-intensive operations, coupled to Matlab (Lin et al., 2005, 2006). Another, a commercial product, is Agilent's Advanced Design System Ptolemy simulator, a time-domain simulator capable of "co-simulating" with Matlab (examples of usage in Braunschvig et al. (2006); Gels et al. (2007)).

Communications-system simulation is completely described in Jeruchim et al. (2000). In this section and in the previous one, we discuss some of the major points.

13.3.1 Pitfalls of Simulation

First some words of serious caution about simulation. Communications-system simulations sometimes drive important or mid-level decisions, for example, that a particular hardware design is satisfactory, that a particular hardware design is better than another, or that a particular parameter being out of spec in the built hardware does not matter in end-to-end system performance. However, relying on simulation results is a risky proposition; the author has been provided more garbage as supposedly final results than valid answers. This happens less often with hardware tests. There are several reasons for this:

- At two companies the author has worked at, people newly out of graduate school were given the job of simulating. At best, they may have studied communications in graduate school so they know the basic concepts but not the fine points. Additionally, almost every new graduate does not understand the importance of reaching the right answer in payload calculations (this also

applied to the author early on). In school getting the theory right is more important than getting the answer numerically correct. Specifically for simulation, the new people do not know yet that they must check every step of their modeling and every result every way they can think of (see below). In contrast, with hardware testing, no one would think of just giving the test-equipment manuals to a new employee who would then, unsupervised and before learning the discipline of validating elements and partial integrations, provide his first measurement results to an engineer to base a decision on.

- The most common mistake of people new to communications simulation is to use filters that are either twice as wide or half as wide as they should be. This is because of the existence of various definitions of bandwidth (RF versus baseband, one-sided versus two-sided—Section 10.2). There are many other possible mistakes involving noise power, SNR versus E_s/N_0, aliasing, sampling rate and sample-run length, and so on.
- Communications-system simulation packages are deceptively easy to use. Ease of use is a selling point for the packages. This is dangerous. Some people without a communications background, who do not even know the basics let alone the fine points, and without the other crucial knowledge of discrete-time-and-frequency signal processing, think they can get valid results because they are able to put together something that runs.
- A communications simulation can be put together and run so fast that even a knowledgeable and experienced user can easily make mistakes. It is easy to forget to check that every element from the last, similar simulation is still appropriate to the current simulation. It is tempting not to take the time to think through the details of the best way to do a new simulation but just to plunge right in. In contrast, hardware testing takes time so there is more opportunity to think.
- Customers of simulation may not have the background to judge whether or not results are valid. (There are ways to protect oneself, though—see below.)

Any customer can be almost sure that simulation results are valid if he takes great care in accepting them. No matter how much experience or education the simulator has, the customer must require the simulator to do the following at each results presentation:

a. Show the overall simulation model in some detail. In an appendix, show every modeling assumption made of every element, including plots as appropriate that are fully labeled.
b. Describe confirmation of the simulation model including every element and partial system integrations. Confirming an element model means confirming that the chosen value of every model parameter yields the expected characteristic.
c. Show exhaustive confirmation of results including comparison against theoretical results for simple but similar cases, journal-article results for

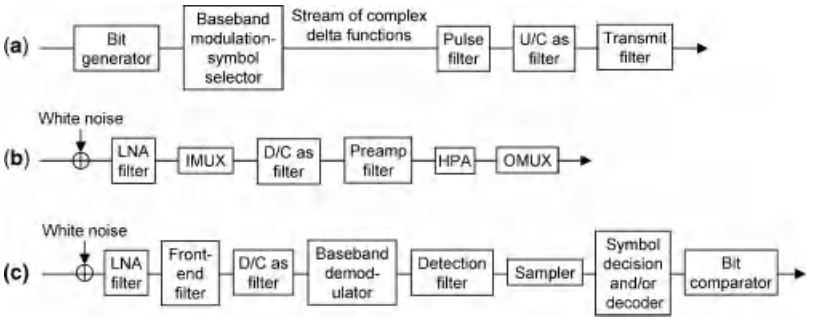


FIGURE 13.6 Typical simulation model of system with bent-pipe payload: (a) ground transmitter, (b) payload, and (c) ground receiver.

one-parameter-variation cases, and if available hardware results and earlier proven simulation results.

d. If the simulator is rather new to it, he must show that he has talked over his model, his model confirmation, and his results with an experienced person.

13.3.2 System Model Specialized to Simulation

Figure 13.6 shows a typical simulation model of a satellite communications system with a bent-pipe payload. The simulation works on the complex baseband equivalent of RF and IF (Section 10.2.1.2). It uses discrete time and frequency. Some operations are carried out in the time domain, such as data-bit stream generation, modulation, HPA action, and from the demodulator through to symbol decision. Filtering is usually done in the frequency domain. Noise is added in either domain or is in some cases handled analytically (see Section 13.3.4). Setting the optimal phase and timing in the ground receiver is done "by hand." In this particular model, residual phase noise and residual timing jitter, where residual is what is left after tracking in the ground receiver, are represented as a part of the white noise added to the ground-receiver front end. Spurious signals that arise from the payload are also represented as a part of that white noise. However, in some circumstances, for example, for a coded system, this is not an appropriate way to model (Chie, 2011).

Additional system elements that must sometimes be modeled are listed above in Section 13.2.1. Other options would be to more explicitly model residual phase error, residual timing jitter, or spurious signals, especially if the payload's contribution is under study (see Section 13.3.5).

13.3.3 When a Signal Distortion Can Be Ignored

Any one source of signal distortion can be ignored if it is small enough. If it can be modeled as additional additive noise, one can decide what error in BER is tolerable and then refer to the appropriate curve among those of Figure 13.4 to see what C/I is

negligible. Those curves are for QPSK; for other modulations a simulation could be performed to yield the similar appropriate curve.

13.3.4 Simulating Additive Noise

For a bent-pipe repeater, the uplink *C*/*N* is set by the payload. Since one of the repeater requirements is on noise figure referred to the receive-antenna terminal (Section 7.12), the equivalent noise level there can be easily calculated. The signal level there can be obtained from the signal level into the payload and the antenna gain. There are two cases, the CAMP in ALC mode and the CAMP in fixed-gain mode (FGM). In the first, the repeater's noise figure depends on the ALC gain, which depends on the input signal level. In the second case, the gain is fixed, so the noise figure is independent of input signal level.

In modeling, additive noise can be used to represent not just thermal noise but also antenna noise and some signal distortions. There are two main ways to model additive noise, besides ignoring it.

One is the Monte Carlo technique, where samples of noise are explicitly generated by a random-number generator and added to the signal samples (Jeruchim et al., 2000).

The difficulty in modeling the additive noise for a system with a bent-pipe payload is the existence of the nonlinear element, the HPA, in the middle of the system. If it were not there, that is, the system were linear, the thermal noise and the other noises could be evaluated at the places they come into the system, added up, and modeled as one additive noise in the ground-receiver front end. Monte Carlo would not be necessary (see just below). However, the HPA is a difficulty. It changes the characteristics of the noise (partially suppressing the radial component when the HPA is near saturation), and the presence of the noise in the HPA changes the way the signal reacts to the HPA. If uplink noise must be modeled, Monte Carlo is the best way to do it.

So to model noise, the first step is to assess the uplink and downlink noise. It is not often the case that the uplink *C*/*N* and the downlink *C*/*N* are about equal. If one is much bigger than the other, it can be assumed ideal. Otherwise, both uplink and downlink noise must be treated.

The good thing about downlink noise is that it can be modeled **semianalytically**—the noise need not be explicitly generated. For every complex-valued symbol phasor out of the ground-receiver's sampler, the SER and/or the BER is calculated, for example, for QPSK by the complementary error function (erfc). The running sum of the SER/BERs is kept and at the end divided by the total number of symbols/bits generated. (Each calculated SER/BER is actually a **conditional** SER/BER, conditioned on the set of nearby symbols. The number of symbols in the set is the system memory length, as in Section 13.2.4.) In the case that both uplink and downlink noise must be modeled, a two-dimensional Monte Carlo simulation, which would be prohibitive, can thus be avoided.

Let us summarize the way to model both uplink and downlink noise. The signal is generated at a sample rate of about 8 or 16 samples per symbol. A stream of

correlated noise samples is generated at the same sample rate and added to the signal samples. The noisy signal samples are put through all the pre-nonlinearity filters. The memoryless baseband nonlinearity performs its action on each noisy sample. The nonlinearity output consists of a signal component, noise, and IMPs. The post-nonlinearity samples pass through any filter at the payload output, the ground-receiver front-end filter, the baseband demodulator, the detection filter, and the symbol-rate sampler. The downlink noise is then treated by the semianalytic method.

13.3.5 Simulating Other Parameters That Vary

To have the system simulation model take into account the variation of one parameter besides noise, the pdf of the parameter must be known or approximated (often by Gaussian). A set of values and weights representing the pdf must be obtained. For example, for a uniform pdf the values could be in the middle of even intervals covering the parameter's range and the weights would equal the inverse of the number of intervals. For each value in the set, the parameter is set to that value and the SER/BER is obtained as described in the previous section. The SER/BER thus obtained is a conditional SER/BER. The conditional SER/BERs are weighted and summed, yielding the SER/BER averaged over the parameter's variation.

A parameter with a Gaussian pdf has its variation best modeled, not by the time-consuming Monte Carlo technique, but by Gauss–Hermite integration (Section 8.5). Some Gaussian-distributed parameters are the following:

- Residual phase error in demodulator
- Residual timing error in symbol-timing recovery
- Antenna pointing error.

The most efficient way to evaluate the SER/BER taking into account the parameter variation is to perform the variation on the symbol-rate samples in the receiver, when possible.

The variation of two or more independent parameters can be taken into account by performing multidimensional conditioning and weighting. If the parameters are not independent, a set of independent parameters must be obtained from which the dependent set can be calculated.

13.3.6 HPA Simulation

If the HPA includes a linearizer, the HPA model is of the combined linearizer and amplifier, that is, the two are not modeled separately.

A memoryless nonlinearity is completely described by P_{out} versus P_{in} and phase shift versus P_{in} curves (Section 6.2.1). The HPA model is this simple if the HPA's small-signal frequency response $H_{\text{ss}}(f)$ and the gain-only saturated-signal frequency response $|H_{\text{sat}}(f)|$ are nearly flat over the spectrum band of the signal of interest. If there is a large interferer, the responses would have to be practically the same over its spectrum band as over the spectrum band of the signal of interest.

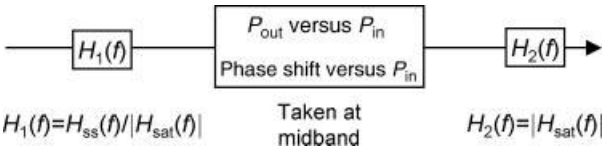


FIGURE 13.7 Three-box model for wide-band HPA (after Silva et al. (2005)).

If the conditions are not satisfied, the recommended HPA model is the **three-box model**, consisting of a filter $H_1(f)$ followed by a memoryless nonlinearity and another filter $H_2(f)$, as illustrated and partially defined in Figure 13.7. There has been much study and experimentation over the years on the best way to represent the HPA, and many models have been proposed (Silva et al., 2005). The three-box model matches the HPA's response at small signal and at large signal but not especially in between. However, the model has big advantages: it is easily characterized (indeed, the necessary measurements are commonly required of the HPA manufacturer by the payload manufacturer), quick-executing, and independent (for good or ill) of the exact signals put through the HPA (Silva et al., 2005). The model is accurate enough for all simulations except possibly when the signals are very wideband (on the order of 15% of the center frequency) or when the simulation's focus is on detailed HPA performance. The only other practical contender is the family of "polyspectral" methods. They are more accurate than the three-box model and quick-executing, but they are applicable only to the signal(s) with which a special time-domain HPA characterization is made (Silva et al., 2005).

There are some details to be accommodated in the implementation of a three-box model. The gain-only frequency response $|H_{sat}(f)|$ is to be measured with a CW power that saturates the HPA at midband. $H_1(f)$ and $H_2(f)$ are to be scaled so that both have unity gain and zero phase at midband (Silva et al., 2005).

Large CW or modulated-signal interferers that go through the HPA with the signal of interest must be explicitly simulated, as they create significant IMPs.

13.3.7 Coding, Decoding, and Interleaving Simulation

Some codes require, for best performance, that the errors in the decoder input be mutually independent. There are at least two situations when the independence may not hold. One is when the decoder input has been subjected to burst errors, for example, from fading or jamming. The other situation is when the decoder input is actually the output of another decoder and this other decoder puts out error bursts, for example, it is a Viterbi decoder of convolutional code.

All decoders require that the rate of input errors not be too large, and what is too large depends on the code and the decoding method.

Upon finding out that either necessity will likely be lacking, a system designer would insert an appropriate **interleaver** and matching deinterleaver into the system. An interleaver mixes up and spreads out bits in a large input extent (see Proakis and

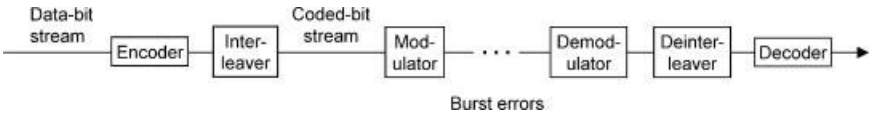


FIGURE 13.8 Coding with interleaver.

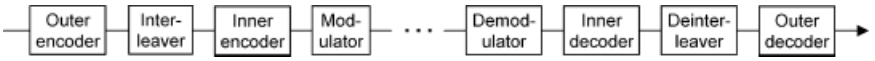


FIGURE 13.9 Concatenated coding with interleaver.

Salehi (2008)). The situation with one encoder/decoder is shown in Figure 13.8 and with two is shown in Figure 13.9.

If errors into the decoder can be assumed independent, coding/decoding need not be explicitly simulated but uncoded BER-to-coded BER curves can be used.

13.3.8 Basic Signal-Processing Considerations

When the simulation alternates between the time and frequency domains, the signal must be defined on discrete, not continuous, points in both domains. The following applies to the signal when it is first considered as defined on continuous time and continuous frequency:

Signal has period $T_{\text{per}} \Leftrightarrow$ Fourier transform is delta functions at multiples of $1/T_{\text{per}}$

Signal is delta functions at multiples of $\Delta t \Leftrightarrow$ Fourier transform has period $1/\Delta t$

Then considered as defined on discrete time and discrete frequency, the delta functions are replaced by regular complex numbers. The relationships among the discrete times and the discrete frequencies are illustrated in Figure 13.10. The signal actually does not exist at intermediate times and frequencies. If interpolation must be done, in the time domain, say, it must be done in a way that does not increase the signal's frequency extent.

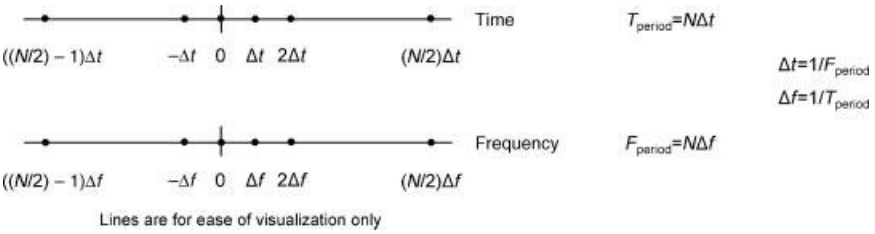


FIGURE 13.10 Relationships among discrete times and discrete frequencies in simulation.

When the simulation alternates between the time and frequency domains, the signal is constructed so as to be periodic in both domains. Usually this means that it goes to zero at the edges. When constructing a signal in the time domain, one must be careful that the signal's extent in the frequency domain would be contained in the available extent, otherwise **aliasing** occurs. In aliasing, the parts of the frequency-domain signal that lie outside the available extent overlap the parts inside the available extent. A similar consideration also applies to constructing a signal in the frequency domain and thinking about its extent in the time domain.

13.4 ADDITIONAL CONSIDERATIONS FOR EMULATION

13.4.1 Emulating Uplink

In emulating the system with hardware, care must be taken to obtain each link's proper C/N_0.

One simple way, sometimes sufficient, to emulate the uplink is shown in Figure 13.11. In the figure, the block marked "attenuators" between the ground transmitter emulator and the bent-pipe payload emulator could just as well be any cascade of passive elements including coaxial cable or waveguide. The cascade has to have the appropriate gain to reduce the signal power to the desired value C_1 into the payload emulator. If the total gain of the passive elements is sufficiently small then the noise temperature into the payload emulator is nearly as low as 290 K. This is fine as long as the T_{10} to be emulated is 290 K. T_{10} is supposed to be the payload's antenna-noise temperature T_{a1}.

When the noise temperature T_{10} into the payload emulator must be larger than 290 K, then a noise source has to be added to the hardware setup, as shown in Figure 13.12. The noise source must provide a noise temperature much higher than 290 K. The required value of the attenuation after the ground transmitter emulator

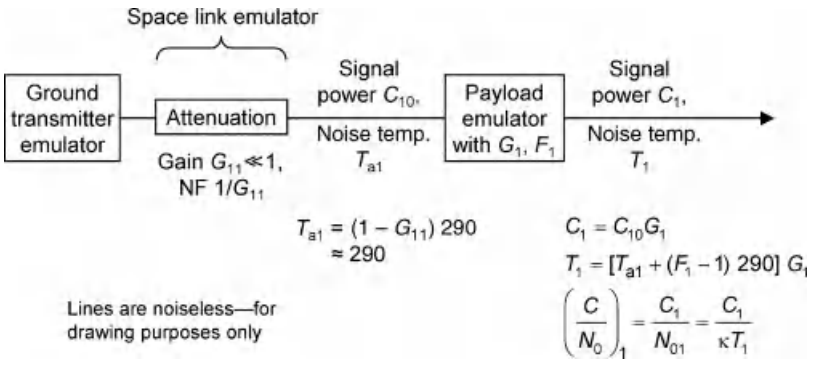


FIGURE 13.11 Uplink C/N_0 emulation for when noise temperature into payload needs to be 290 K, for bent-pipe payload.

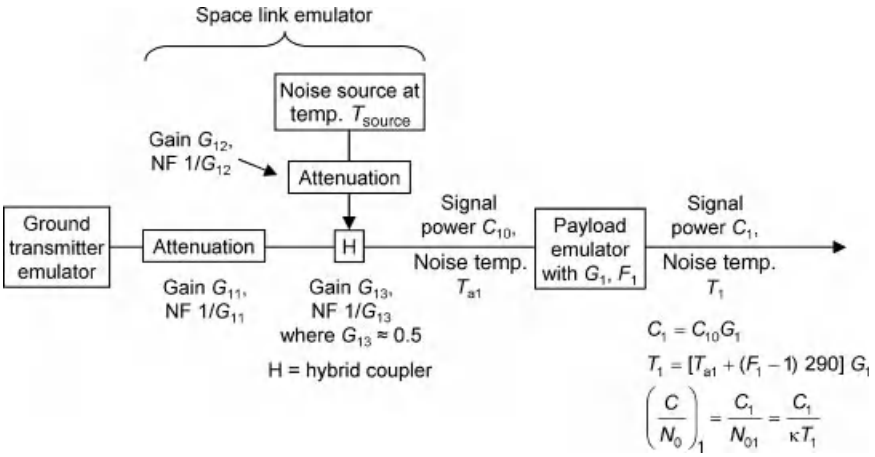


FIGURE 13.12 Uplink C/N_0 emulation for when noise temperature into payload needs to be greater than 290 K, for bent-pipe payload.

is easily found. The attenuator "gain" between the noise source and the hybrid (i.e., 3-dB) coupler is as follows:

$$G_{12} = \frac{T_{10} + 290 - (1 - G_{11})G_{13}290}{(T_{\text{source}} - 290)G_{13}}$$

For a regenerative payload, the uplink emulation is similar except that the payload emulator is only of the payload's front end, as shown in Figure 13.13.

13.4.2 Emulating Downlink

Figure 13.14 shows the downlink emulator for any kind of payload. The issue of antenna noise for the space link emulator is similar to that on the uplink.

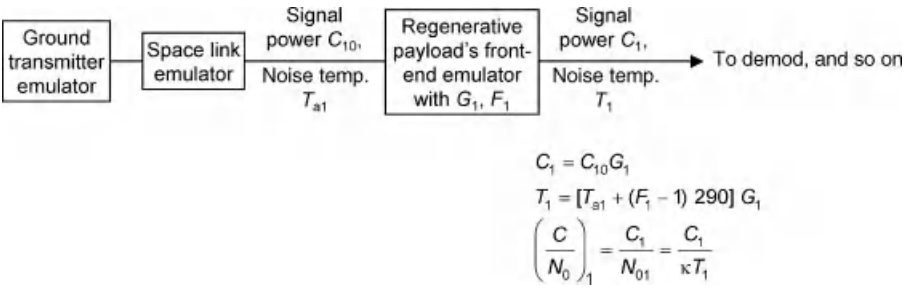


FIGURE 13.13 Uplink C/N_0 emulation for regenerative payload.

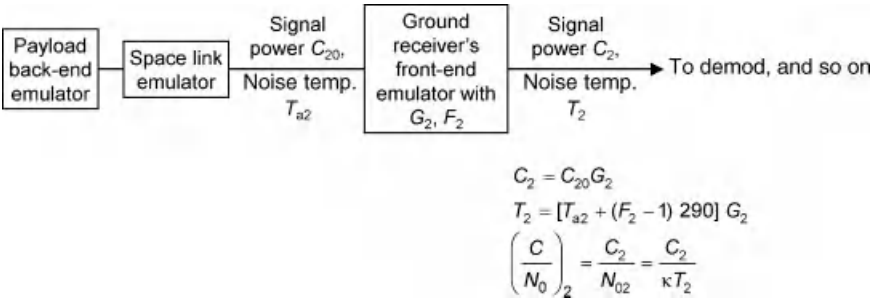


FIGURE 13.14 Downlink C/N_0 emulation for both bent-pipe and regenerative payloads.

Figure 13.15 shows the composite up- and downlink emulation for a bent-pipe payload and confirms that the end-to-end C/N_0 comes out to its proper value (cf. Section 11.2).

13.4.3 Matching Gain Tilt and Parabolic Phase

In emulators of the ground transmitter, the payload, and the ground receiver's front end, for convenience the lengths of waveguide and coax are usually not the same as in the actual hardware. Over a frequency band, both waveguide and coax introduce gain tilt and perhaps parabolic gain distortion, and waveguide introduces parabolic

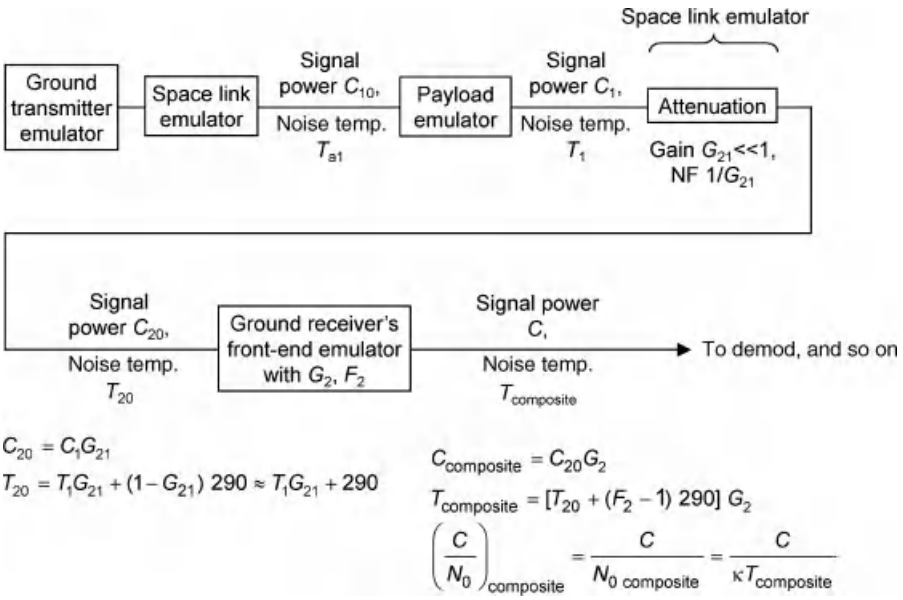


FIGURE 13.15 System C/N_0 emulation for bent-pipe payload.

phase distortion (Sections 4.3.1.2 and 4.3.2.2). These distortions degrade the communications performance (Section 10.4). The lengths of the waveguide or coax used to emulate the space links can potentially make up for some of the difference. In any case, the effects must be assessed. If necessary, a gain and/or phase equalizer can be added to a space-link emulator to better mimic the actual system.

In the actual system, the gain tilt from one antenna exactly compensates the gain tilt from space loss. The gain tilt from the other antenna needs to be emulated.

REFERENCES

Braun WR and McKenzie TM (1984). CLASS: a comprehensive satellite link simulation package. *IEEE Journal on Selected Areas in Communications;* 2 (Jan.); 129–137.

Braunschvig E, Casini E, and Angeletti P (2006). Co-channel signal power measurement methodology in a communication and payload joint simulator. *AIAA International Communications Satellite Systems Conference;* June 11–14; 1–14.

Chie CM, retired from Boeing Satellite Center (2011). Private communication. Oct. 30.

Gels B, Andrews M, Hendry D (2007). Simulation of the effects of Q-band amplifier nonlinearities on non-constant envelope SATCOM waveforms. *Proceedings of IEEE Military Communications Conference;* Oct. 29–31; 1–6.

Jeruchim MC, Balaban P, and Shanmugan KS (2000). *Simulation of Communications Systems: Modeling, Methodology, and Techniques*, 2nd ed. New York: Kluwer Academic/Plenum Publishers.

Lin VS, Speelman RJ, Daniels CI, Grayver E, and Dafesh PA (2005). Hardware accelerated simulation tool (HAST). *IEEE Aerospace Conference;* Mar. 5–12; 1475–1483.

Lin VS, Arredondo A, and Hsu J (2006). Efficient modeling and simulation of nonlinear amplifiers. *IEEE Aerospace Conference;* Mar. 4–11; 1–9.

Papoulis A (1984). *Probability, Random Variables, and Stochastic Processes,* International Student ed., 2nd ed. Singapore: McGraw-Hill, Inc.

Proakis JG and Salehi M (2008). *Digital Communications*, 5th ed., International ed. New York: McGraw-Hill.

Silva CP, Clark CJ, Moulthrop AA, and Muha MS (2005). Survey of characterization techniques for nonlinear communication components and systems. *IEEE Aerospace Conference;* Mar. 5–12; 1713–1737.

INDEX

Satellite Communications Payload and System, First Edition. Teresa M. Braun.