

SECOND EDITION

ELECTRIC POWER SYSTEMS

A CONCEPTUAL INTRODUCTION

ALEXANDRA von MEIER



 IEEE Press

WILEY

Electric Power Systems

IEEE Press
445 Hoes Lane
Piscataway, NJ 08854

IEEE Press Editorial Board
Sarah Spurgeon, *Editor-in-Chief*

Moeness Amin
Jón Atli Benediktsson
Adam Drobot
James Duncan

Ekram Hossain
Brian Johnson
Hai Li
James Lyke
Joydeep Mitra

Desineni Subbaram Naidu
Tony Q. S. Quek
Behzad Razavi
Thomas Robertazzi
Diomidis Spinellis

Electric Power Systems

A Conceptual Introduction

Second Edition

Alexandra von Meier



Copyright © 2024 by The Institute of Electrical and Electronics Engineers, Inc. All rights reserved.

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

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, electronic, mechanical, photocopying, recording, scanning, or otherwise, except as permitted under Section 107 or 108 of the 1976 United States Copyright Act, without either the prior written permission of the Publisher, or authorization through payment of the appropriate per-copy fee to the Copyright Clearance Center, Inc., 222 Rosewood Drive, Danvers, MA 01923, (978) 750-8400, fax (978) 750-4470, or on the web at www.copyright.com. Requests to the Publisher for permission should be addressed to the Permissions Department, John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, (201) 748-6011, fax (201) 748-6008, or online at <http://www.wiley.com/go/permission>.

Trademarks: Wiley and the Wiley logo are trademarks or registered trademarks of John Wiley & Sons, Inc. and/or its affiliates in the United States and other countries and may not be used without written permission. All other trademarks are the property of their respective owners. John Wiley & Sons, Inc. is not associated with any product or vendor mentioned in this book.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representations or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages. Further, readers should be aware that websites listed in this work may have changed or disappeared between when this work was written and when it is read. Neither the publisher nor authors shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Library of Congress Cataloging-in-Publication Data

Names: Meier, Alexandra von, author.

Title: Electric power systems : a conceptual introduction / Alexandra von Meier.

Description: Second edition. | Hoboken, New Jersey : John Wiley & Sons, Inc., [2024] | Includes index.

Identifiers: LCCN 2024011469 (print) | LCCN 2024011470 (ebook) | ISBN 9781394241002 (Hardback) | ISBN 9781394241026 (adobe pdf) | ISBN 9781394241019 (epub)

Subjects: LCSH: Electric power systems.

Classification: LCC TK1005 .M37 2024 (print) | LCC TK1005 (ebook) | DDC 621.31-dc23/eng/20240402

LC record available at <https://lccn.loc.gov/2024011469>

LC ebook record available at <https://lccn.loc.gov/2024011470>

Cover Design: Wiley

Cover Image: Courtesy of the Author

Set in 9.5/12.5pt STIXTwoText by Straive, Chennai, India

*To my late grandfather
Karl Wilhelm Clauberg
who introduced me to
The Joy of Explaining Things*

Contents

List of Figures xvii

Preface xxv

Acknowledgments xxix

About the Companion Website xxxi

1	Physics of Electricity	1
1.1	Basic Quantities	1
1.1.1	Charge	1
1.1.2	Potential or Voltage	2
1.1.3	Ground	4
1.1.4	Conductivity	4
1.1.5	Current	5
1.2	Ohm's Law	7
1.2.1	Resistance	7
1.2.2	Conductance	9
1.2.3	Insulation	10
1.3	Circuit Fundamentals	10
1.3.1	Static Charge	10
1.3.2	Closing a Circuit	11
1.3.3	Voltage Drop	11
1.3.4	Electric Shock	12
1.4	Resistive Heating	13
1.4.1	Calculating Resistive Heating	13
1.4.2	Transmission Voltage and Resistive Losses	16
1.5	Electric and Magnetic Fields	17
1.5.1	The Field as a Concept	17
1.5.2	Electric Fields	18
1.5.3	Magnetic Fields	19
1.5.4	Electromagnetic Induction	21
1.5.5	Electromagnetic Fields and Health Effects	23
1.5.6	Electromagnetic Radiation	23
	Problems and Questions	26

2	DC Circuit Analysis	29
2.1	Modeling Circuits	29
2.2	Series and Parallel Circuits	30
2.2.1	Resistance in Series	31
2.2.2	Resistance in Parallel	31
2.2.3	Network Reduction	33
2.2.4	Dual Concepts	34
2.2.5	Practical Aspects	34
2.3	Kirchhoff's Laws	35
2.3.1	Kirchhoff's Voltage Law	36
2.3.2	Kirchhoff's Current Law	36
2.3.3	Application to Simple Circuits	37
2.4	The Superposition Principle	39
2.5	Thévenin and Norton Equivalent Circuits	41
2.5.1	One-ports: Battery and PV Cell	41
2.5.2	Thévenin and Norton Theorems	44
2.6	Magnetic Circuits	48
	Problems and Questions	51
3	AC Power	55
3.1	Alternating Current and Voltage	55
3.1.1	Historical Notes	55
3.1.2	Mathematical Description of Alternating Current	56
3.1.3	The rms Value	59
3.2	Power for the Resistive Case	60
3.2.1	Power Dissipated Versus Transmitted	61
3.2.2	Time-Varying Resistive Power	62
3.3	Impedance	63
3.3.1	Inductance	63
3.3.2	Inductive Reactance	67
3.3.3	Capacitance	67
3.3.4	Capacitive Reactance	69
3.3.5	Complex Numbers	71
3.3.6	Complex Impedance	73
3.3.7	Complex Admittance	75
3.4	Complex Power	77
3.4.1	Real Power and Power Factor	78
3.4.2	Reactive Power	79
3.4.3	Power in the Complex Plane	81
3.4.4	Reactive Power in the Power System Context	83
3.4.5	Reactive Compensation	84
3.5	Phasors	86
3.5.1	Introduction	86
3.5.2	Derivation	87
3.5.3	Euler's Equation	89

3.5.4	Operations with Phasors	91
3.5.5	Ohm's Law in Complex Form	91
3.5.6	Kirchhoff's Laws with Phasors	92
3.5.7	Complex Power in Phasor Notation	94
	Problems and Questions	96
4	Three-Phase Power	101
4.1	Three-Phase Basics	101
4.1.1	Rationale for Three Phases	101
4.1.2	Number of Phases	104
4.1.3	Balancing Loads	105
4.1.4	Delta and Wye Connections	106
4.1.5	Practical Aspects	109
4.1.6	Three-phase Complex Power	110
4.1.7	Three-phase Impedance	111
4.2	Symmetrical Components	111
4.2.1	Converting Symmetrical Components	114
4.2.2	Ohm's Law with Symmetrical Components	116
4.3	Direct and Quadrature Components	117
	Problems and Questions	117
5	Power Quality	121
5.1	Voltage	121
5.1.1	Conservation Voltage Reduction	123
5.2	Frequency	125
5.3	Waveform and Harmonics	126
5.3.1	Current Versus Voltage Harmonics	128
5.3.2	Quantifying Harmonic Distortion	130
5.3.3	Distortion Power Factor	131
5.3.4	Transformers and Triplen Harmonics	132
	Problems and Questions	133
6	Loads	135
6.1	Types of Loads	135
6.1.1	Resistive Loads	136
6.1.2	Dimmer Circuits	137
6.1.3	Motors	139
6.1.4	Electronic Devices	142
6.1.5	Electric Vehicles	143
6.2	Single- and Multiphase Connections	144
6.3	Voltage Response of Loads	146
6.3.1	ZIP Load Model	146
6.3.2	Transient Response	149
6.4	Load in Aggregate	150
6.4.1	Historical Context	150

6.4.2	Coincident and Noncoincident Demand	151
6.4.3	Load Profiles and Load Duration Curve	151
6.4.4	Managing Load	154
	Problems and Questions	156
7	Transmission and Distribution Systems	159
7.1	System Structure	159
7.1.1	Interconnection	159
7.1.2	Structural Features	163
7.1.3	International Differences in Distribution System Design	165
7.1.4	Stations and Substations	166
7.1.5	Topology	167
7.1.6	Power Islands	170
7.1.7	Loop Flow	171
7.1.8	Reconfiguring the System	173
7.2	Qualitative Characteristics of Power Lines	174
7.2.1	Conductors	174
7.2.2	Bundled Conductors	176
7.2.3	Towers, Insulators, and Other Components	177
7.2.4	DC Transmission	178
7.2.5	Superconducting Transmission	181
7.3	Loading	182
7.3.1	Thermal Limits	182
7.3.2	Stability Limit	183
7.3.3	Surge Impedance Loading	184
7.4	Voltage Control	185
7.4.1	Tap Changers	187
7.4.2	Reactive Compensation	188
7.5	Protection	190
7.5.1	Basics of Protection and Protective Devices	190
7.5.2	Protection Coordination	194
7.5.3	Unsymmetrical and Asymmetrical	197
	Problems and Questions	201
8	Transformers	203
8.1	General Properties	203
8.2	Transformer Heating	205
8.3	Delta and Wye Transformers	206
8.4	Autotransformers	208
8.5	Transformer Modeling	210
8.5.1	Nonideal Characteristics	210
8.5.2	Referred Impedance	213
8.5.3	Open-Circuit and Short-Circuit Tests	215
8.6	Voltage Regulation	216
8.6.1	Approximation	217

8.7	Per-unit System	218
	Problems and Questions	222
9	Analyzing Transmission Lines	225
9.1	Transmission Line Inductance	225
9.1.1	Internal Flux Linkage	227
9.1.2	External Flux Linkage	229
9.1.3	Per-Phase Inductance	230
9.1.4	Geometric Mean Distance and Radius	231
9.2	Transmission Line Capacitance	234
9.2.1.1	Ground Effects	237
9.3	ABCD Parameters	238
9.3.1	Two-Ports	238
9.3.2	Line Models Overview	239
9.3.3	Short Line Model	240
9.3.4	Short Line Phasor Relationship	241
9.3.5	Medium Line Model	242
9.3.5.1	Charging Current	245
9.3.6	Medium Line Qualitative Observations	245
9.3.7	Long Line Model: Introduction	246
9.3.8	Long Line Model: Wave Behavior	246
9.3.9	Long Line Model: ABCD Parameters	250
9.3.9.1	Lumped-Circuit Equivalent	251
9.3.10	Lossless Line	252
	Problems and Questions	253
10	Machines	257
10.1	The Simple Generator	258
10.2	D.C. Machine	261
10.2.1	The Paper Clip Motor	262
10.3	The Synchronous Generator	264
10.3.1	Basic Components and Functioning	264
10.3.2	Number of Poles	267
10.3.3	Other Design Aspects	269
10.4	Operational Control	270
10.4.1	Single Generator: Real Power	270
10.4.2	Single Generator: Reactive Power	272
10.4.3	Multiple Generators: Real Power	277
10.4.4	Multiple Generators: Reactive Power	280
10.5	Operating Limits	283
10.6	The Induction Machine	285
10.6.1	General Characteristics	285
10.6.2	Electromagnetic Characteristics	287
10.6.3	Reluctance Machine	289
10.7	Modeling Generators	291

10.7.1	Equivalent Circuit Model	293
10.7.2	Over- and Underexcitation	295
10.7.3	Power Transfer	295
	Problems and Questions	297
11	Matching Generation and Load	299
11.1	Load Frequency Control	299
11.1.1	Inertia	300
11.1.2	Primary Frequency Regulation	301
11.1.3	Secondary Frequency Regulation	303
11.1.3.1	Multiple Generators	304
11.1.4	Frequency Tolerance	306
11.1.5	Area Control Error	307
11.2	Economic Dispatch	312
11.2.1	Filling in the Load Duration Curve	312
11.2.2	Lagrangian Method	314
	Problems and Questions	318
12	Power Flow	321
12.1	Introduction	321
12.2	The Power Flow Problem	322
12.2.1	Network Representation	322
12.2.2	Choice of Variables	323
12.2.3	Nonlinearity	325
12.2.4	Types of Buses	327
12.2.5	Variables for Balancing Real Power	327
12.2.6	Variables for Balancing Reactive Power	328
12.2.7	The Slack Bus	329
12.2.8	Summary of Variables	331
12.3	Example with Interpretation of Results	331
12.3.1	Six-bus Example	331
12.3.2	Tweaking the Case	334
12.3.3	Conceptualizing Power Flow	336
12.4	Power Flow Equations and Solution Methods	339
12.4.1	Derivation of Power Flow Equations	339
12.4.2	The Bus Admittance Matrix	341
12.4.3	Solution Methods	345
12.4.4	Iterative Computation	346
12.4.5	Power Flow Example	350
12.4.5.1	Low-voltage Solution	353
12.4.6	Shortcuts	355
12.4.6.1	Dishonest Newton–Raphson	355
12.4.6.2	Decoupled Power Flow	355
12.4.6.3	Fast-Decoupled Power Flow	356
12.4.6.4	DC Power Flow	357

12.5	Applications	360
12.5.1	Optimal Power Flow	360
12.5.2	State Estimation	361
12.6	LinDistFlow	363
12.6.1	Derivation	366
	Problems and Questions	367
13	Limits	369
13.1	Adequacy	369
13.2	Reliability	370
13.2.1	Measures of Reliability	370
13.2.2	Valuing Reliability	372
13.3	Security	374
13.4	Stability	376
13.4.1	Overview	376
13.4.2	The Concept of Stability	377
13.4.3	Angle Stability	378
13.4.4	Transient Angle Stability	382
13.4.5	Voltage Stability	392
13.5	Power Transfer Limits	394
13.5.1	<i>P–V</i> Curve	397
13.5.2	<i>V–Q</i> Curve	400
	Problems and Questions	403
14	Power Electronics	405
14.1	Power Conversion: Introduction	405
14.2	Legacy Power Conversion Technologies	406
14.2.1	Mercury Arc Valves	406
14.2.2	Vacuum Diodes and Triodes	408
14.3	Solid-State Technology	408
14.3.1	p–n Junctions and Diodes	409
14.3.2	Transistors	411
14.3.3	Thyristors	413
14.4	Inverters	415
14.4.1	Basic Inverter Function	415
14.4.2	Sample Inverter Circuit	418
14.4.3	Inverter Control	419
14.5	FACTS	423
15	Resources	425
15.1	Generation Resources	425
15.1.1	Hydroelectricity	425
15.1.2	Thermal Generation	426
15.1.2.1	Fossil Fuels	429
15.1.2.2	Biomass	429

15.1.2.3	Geothermal Power	430
15.1.2.4	Nuclear Power	430
15.1.2.5	Concentrating Solar Power	432
15.1.3	Solar Photovoltaics	433
15.1.4	Wind Power	435
15.2	Distributed Generation	437
15.2.1	DG Resources	437
15.2.1.1	Fuel Cells	437
15.2.1.2	Microturbines	438
15.2.1.3	Small Generators	438
15.2.1.4	Small Wind Turbines	439
15.2.2	DG Integration	439
15.3	Storage	443
15.3.1	Hydroelectric Storage	443
15.3.2	Batteries	444
15.3.3	Other Storage Technologies	447
15.3.3.1	Thermal Storage	447
15.3.3.2	Compressed Air	447
15.3.3.3	Flywheels	447
15.3.3.4	SMES	448
15.3.3.5	Supercapacitors	448
15.3.3.6	Hydrogen	448
15.4	Microgrids	449
16	Making the System Work	453
16.1	Time Scales for Operation and Control	454
16.1.1	Fast Response	455
16.1.2	Real-Time Operation	456
16.1.3	Scheduling	457
16.1.4	Planning	459
16.2	Measurement and Data	460
16.2.1	Historical Notes	460
16.2.2	Physical Measurements	462
16.2.3	Reporting Measurements	463
16.2.4	PMUs	465
16.2.5	Actionable Intelligence and Automation	468
16.3	Human Factors	469
16.3.1	Operators and Engineers	469
16.3.2	Cognitive Representations of Power Systems	470
16.3.3	Operational Criteria	473
16.3.4	Implications for Technological Innovation	477
16.4	Strategic Perspectives	479

- 16.4.1 Decarbonization 479
16.4.2 Markets 480

Appendix A Symbols, Units, Abbreviations, and Acronyms 487

Index 493

List of Figures

- Figure 1.1** Electric field of (a) a single charge and (b) two opposite charges. 18
- Figure 1.2** Magnetic field around a current-carrying wire. 20
- Figure 1.3** The Lorentz force acting on charges moving through a magnetic field. 22
- Figure 1.4** An electromagnetic wave. 24
- Figure 2.1** Resistors in series. 31
- Figure 2.2** Resistors in parallel. 32
- Figure 2.3** Network reduction. 33
- Figure 2.4** Kirchhoff's voltage law. 36
- Figure 2.5** Kirchhoff's current law. 37
- Figure 2.6** Superposition. (a) Full circuit. (b) Circuit with only the voltage source. (c) Circuit with only the current source. 40
- Figure 2.7** A one-port representing a relationship between voltage v and current i , governed by whatever is inside the box. 41
- Figure 2.8** A simple, linear battery operating characteristic (a) and Thévenin equivalent circuit (b). 43
- Figure 2.9** (a) The current–voltage operating characteristic or I – V curve for an ideal photovoltaic cell, and (b) the circuit model that reproduces this behavior using an ideal current source and diode. 44
- Figure 2.10** Operating characteristic for a linear one-port with Thévenin and Norton equivalents. 46
- Figure 2.11** Sample linear circuit. 46
- Figure 2.12** Thévenin (a) and Norton (b) equivalents of the circuit in Figure 2.11. 47
- Figure 2.13** A magnetic circuit. 49
- Figure 2.14** Network of resistors. 51
- Figure 3.1** A sine function $f(t) = A \sin \omega t$ plotted against time t or, equivalently, angle ωt . 57
- Figure 3.2** Sinusoidal alternating current with phase shift. 57
- Figure 3.3** Derivation of the rms value. 60
- Figure 3.4** A basic inductor, or solenoid. 64
- Figure 3.5** Current lagging voltage by 90° . 65

- Figure 3.6** A basic capacitor, with arrows indicating the electric field. 67
- Figure 3.7** Current leading voltage by 90° . 69
- Figure 3.8** The number $C = 3 + j4$ in the complex plane. In polar notation, $C = 5\angle 53.1^\circ$. 72
- Figure 3.9** The impedance Z represented in the complex plane, with resistance R in the real direction and reactance jX in the imaginary direction. 74
- Figure 3.10** Simple series circuit to illustrate the addition of impedances. 76
- Figure 3.11** Parallel circuit to illustrate the addition of admittances, using the same elements as in Figure 3.10. 77
- Figure 3.12** Power as the product of voltage and current, with current lagging behind voltage by a phase angle difference θ . 78
- Figure 3.13** Complex power S with real power P in the real and reactive power Q in the imaginary direction. 82
- Figure 3.14** A phasor in the complex plane. 87
- Figure 3.15** Phasors in relation to each other. 88
- Figure 3.16** Complex exponentials. An increasing imaginary exponent corresponds to counterclockwise rotation in the complex plane. 90
- Figure 3.17** Series circuit to illustrate KVL for the complex case. 92
- Figure 3.18** Phasor addition of voltages $V = V_R + V_L + V_C$ associated with Figure 3.17 in the complex plane. 93
- Figure 3.19** Parallel circuit to illustrate KCL for the complex case. 93
- Figure 3.20** Phasor addition in the complex plane of currents $I = I_R + I_L + I_C$ associated with Figure 3.19. 93
- Figure 3.21** Two voltage sources that could be acting as generators or loads. 95
- Figure 4.1** Three balanced single-phase a.c. currents. 102
- Figure 4.2** Three phases with and without neutral return; (a) three phases with common return; (b) three phases with neutral removed. 103
- Figure 4.3** (a) Delta and (b) wye connections. 107
- Figure 4.4** Phase-to-phase (line-to-line) voltage, seen in the time domain as the difference between a pair of phase-to-neutral (line-to-neutral) voltages. 107
- Figure 4.5** Line-to-line and line-to-neutral voltages represented in the phasor domain. 108
- Figure 4.6** Currents in a wye connection. 108
- Figure 4.7** Currents in a delta connection. 109
- Figure 4.8** Current phasors showing the relationships between line currents (I_a , I_b , and I_c) and the currents through a delta-connected load (I_{ab} , I_{bc} , and I_{ca}). 109
- Figure 4.9** Positive-, negative-, and zero-sequence components. Each trio includes three phasors of equal magnitude, imagined as rotating counterclockwise. 112
- Figure 4.10** Vector addition of the symmetrical components in Figure 4.10, corresponding to Eq. 4.6. 113
- Figure 4.11** Various international practices for delta connections. Source: A. McEachern, ([2002] IEEE). 118
- Figure 5.1** ITIC curve. Source: public domain. 124

- Figure 5.2** Transient voltage waveform disturbance, as seen with a PQube power quality recorder. Source: Courtesy of Alex McEachern. 127
- Figure 5.3** Simulated current drawn by a highly nonlinear load and its effect on voltage due to source impedance, visualized in the Power Quality Teaching Toy. 129
- Figure 5.4** The third harmonic of all three phases coincides. 133
- Figure 6.1** Standard U.S. electrical outlet or wall socket. 144
- Figure 6.2** Transformer taps and multiphase service. 145
- Figure 6.3** Example of a load profile from California ISO. Source: Adapted from California Independent System Operator. 152
- Figure 6.4** Example of a load duration curve. Source: Adapted from California Energy Commission/State of California. 152
- Figure 6.5** Load profile on a mild, sunny day. Source: Adapted from California Independent System Operator. 153
- Figure 7.1** Historical growth of generation unit size and transmission voltage. Source: Adapted from Economic Regulatory Administration, 1981. 160
- Figure 7.2** Regions and interconnections of the U.S. electric grid. Source: U.S. Energy Information Agency, 2016/Public Domain. 162
- Figure 7.3** Real-time frequency in the major synchronous interconnections in North America, as seen on FNET (University of Tennessee, Knoxville and Oak Ridge National Laboratory). 163
- Figure 7.4** One-line diagram showing basic power system structure. 164
- Figure 7.5** North American (a) and European (b) distribution systems. 166
- Figure 7.6** Distribution substation layout. 167
- Figure 7.7** Distribution substation. Source: Courtesy of Marshal F. Merriam. 168
- Figure 7.8** Radial distribution system. 168
- Figure 7.9** Loop system. 169
- Figure 7.10** Networked topology. 170
- Figure 7.11** Loop flow. 172
- Figure 7.12** Sample transmission-line dimensions. Source: From EPRI, 1977. 175
- Figure 7.13** Transposition tower for a 380-kV line in Germany, carrying two three-phase circuits. 179
- Figure 7.14** The ±500 kV Pacific DC Intertie (PDCI) and a three-phase 230 kV a.c. transmission line near Bishop, California. 180
- Figure 7.15** Real power flow as a function of voltage phase angle difference. 184
- Figure 7.16** Thermal and stability limits for a hypothetical line. 184
- Figure 7.17** Voltage drop along a distribution feeder. 186
- Figure 7.18** Effect of a line voltage regulator on the voltage profile of a radial distribution feeder. 187
- Figure 7.19** Effect of a capacitor on the voltage profile of a radial distribution feeder. 189
- Figure 7.20** Sample time-current characteristic of a relay. 191

- Figure 7.21** Recloser operation with transient fault. 193
- Figure 7.22** Recloser operation with a permanent fault. 194
- Figure 7.23** Example of protection zones with oil circuit breakers (OCB) and recloser (REC). 194
- Figure 7.24** Protection zones—fuses. 195
- Figure 7.25** Protection zones—reclosers. 195
- Figure 7.26** Protection zones—relay. 196
- Figure 7.27** Sample coordination with time–current curves for a fuse and a recloser. 196
- Figure 7.28** R - L circuit of fault loop shorted at $t = 0$. 198
- Figure 7.29** The role of fault inception: current increases during the entire time that voltage remains positive on the first cycle. 199
- Figure 7.30** Symmetrical (a) and asymmetrical (b) current with a decaying d.c. component due to the timing of the fault’s inception. 199
- Figure 8.1** Transformer concept. A realistic transformer performs better if the core forms a closed loop, as shown in Figure 2.13. 204
- Figure 8.2** Transformer at a distribution substation. Source: Courtesy of Pacific Gas & Electric. 207
- Figure 8.3** Three-phase transformer connections. 208
- Figure 8.4** Autotransformer. 209
- Figure 8.5** Magnetization versus magnetic field strength for a theoretical transformer core, illustrating hysteresis. 212
- Figure 8.6** Equivalent circuit model for a realistic transformer. 213
- Figure 8.7** Transformer example with series impedance only, neglecting shunt admittance. 214
- Figure 8.8** Transformer example with series elements only (neglecting shunt admittance). 214
- Figure 8.9** Vector addition for voltage drop across a series impedance, in the case where load current is aligned with voltage (p.f. = 1.0). 217
- Figure 8.10** Projecting voltage drop onto the real axis to approximate $|V_{NL}| - |V_{FL}|$ as in Eq. (8.6), for lagging and leading current. 218
- Figure 8.11** Adding series impedances in per-unit across multiple zones with different base voltages. 220
- Figure 9.1** The magnetic field around a single current-carrying wire. 226
- Figure 9.2** Integrating over the magnetic flux density around a single current-carrying wire. 227
- Figure 9.3** Accounting for the fraction of enclosed current at $r < R$ inside the conductor, assuming uniform current density. 228
- Figure 9.4** When the return conductor is inside the path integral, the enclosed current and therefore the magnetic flux is zero. The same applies to a three-phase circuit. 230

- Figure 9.5** Three-phase conductors in an equilateral arrangement, with equal spacing between each. 232
- Figure 9.6** Sample three-phase transmission line geometry. 234
- Figure 9.7** Visualizing capacitance between a pair of conductors. 235
- Figure 9.8** Series addition of capacitance. 236
- Figure 9.9** Short transmission line model, which accounts for only the series impedance $Z = R + jX_L$. This model requires $I_S = I_R$. 240
- Figure 9.10** Voltage drop across a short transmission line for a load of lagging, unity, and leading power factor. 241
- Figure 9.11** Equivalent π -circuit for a medium-length transmission line. 242
- Figure 9.12** Derivation of the interdependence of voltage and current as a function of distance x along a transmission line, with distributed line parameters. 247
- Figure 9.13** (a) Adjacent conductor strands. (b) Conductor strands separated by a distance d within a bundled conductor. 253
- Figure 9.14** Cross section of a three-phase transmission line with bundled conductors. 254
- Figure 10.1** The notional “simple generator.” 259
- Figure 10.2** Changing flux and *emf* versus time. 260
- Figure 10.3** Simple generator with armature reaction. 261
- Figure 10.4** Schematic of a direct-current motor. Source: Unknown. 262
- Figure 10.5** Paperclip motor. 263
- Figure 10.6** Cylindrical rotor and its magnetic field. 265
- Figure 10.7** Schematic arrangement of three-phase stator winding. 266
- Figure 10.8** Contribution to armature reaction from one phase. 266
- Figure 10.9** Three-phase armature reaction. 267
- Figure 10.10** Magnetic field of a four-pole rotor. 268
- Figure 10.11** Controlling generator output with the governor valve. 271
- Figure 10.12** Flux, armature voltage, and current versus time; unity power factor. 273
- Figure 10.13** Flux, armature voltage, and current versus time; lagging power factor. 273
- Figure 10.14** Flux, armature voltage, and current versus time; leading power factor. 274
- Figure 10.15** Rotor and stator field, unity power factor. 274
- Figure 10.16** Rotor and stator field, lagging power factor. 275
- Figure 10.17** Rotor and stator field, leading power factor. 275
- Figure 10.18** Decomposition of the stator field. 275
- Figure 10.19** The power angle. 278
- Figure 10.20** “Difference voltage” due to phase angle shift, resulting in a circulating current in phase with voltage that affects real power exchange. 279
- Figure 10.21** Circulating current with Unit 1 providing more reactive power. 281
- Figure 10.22** Sample allocation of reactive power among generators. 282
- Figure 10.23** Reactive capability curve. 284

- Figure 10.24** Torque versus slip for an induction machine. 286
- Figure 10.25** Illustration from Nikola Tesla's a.c. motor patent. 290
- Figure 10.26** Simplified single-phase round rotor machine, with notional magnetic flux lines from the armature windings. 292
- Figure 10.27** Simple equivalent generator circuit model. 294
- Figure 10.28** Phasor diagram for the equivalent circuit in Figure 10.27. 294
- Figure 11.1** James Watt governor. Source: Routledge/Wikipedia/CC BY 4.0. 301
- Figure 11.2** Droop curve showing the relationship between Δf and ΔP . 302
- Figure 11.3** Droop control with a single generator. 303
- Figure 11.4** Droop control with two generators. Points a are the initial state, b after primary response, and c the final state after secondary response, in which only Unit 2 participates. 304
- Figure 11.5** Visualizing tie line flow between neighboring control areas. 307
- Figure 11.6** Three areas sharing frequency regulation resources (initial operating condition in text example). 309
- Figure 11.7** Grid frequency on a typical day in California, measured at a wall outlet with a micro-phasor measurement unit (PMU) and visualized in the Berkeley Tree Database (BTrDB). 311
- Figure 11.8** Generation scheduling with the load duration curve. 313
- Figure 11.9** Droop curves for two generators. 319
- Figure 11.10** Frequency volatility (a) and nadir for loss-of-generation events (b) in West Berlin, before and after interconnection with Western Europe. Source: unknown. 320
- Figure 12.1** One-line diagram for a fictitious power system. 324
- Figure 12.2** Current I_3 is found by adding currents $I_1 + I_2$, superimposing the two partial circuits on the center branch. 326
- Figure 12.3** In a nonlinear system, we cannot directly solve for V_3 or I_3 . 326
- Figure 12.4** Six-bus power flow example. 332
- Figure 12.5** Six-bus power flow example with losses. 334
- Figure 12.6** Modified six-bus power flow example. 335
- Figure 12.7** Common rotating line shaft in a yarn spinning factory (Leipzig, ca. 1925) as a mechanical analog to a.c. power transfer. Source: Atelier Hermann Walter/Wikimedia Commons/Public Domain. 337
- Figure 12.8** Simple network to illustrate obtaining branch admittances. 343
- Figure 12.9** Simple two-bus network to illustrate allocation of transmission line capacitance. 344
- Figure 12.10** Newton's method. 347
- Figure 12.11** Iterative process of approximating $f(x) = 0$. 348
- Figure 12.12** Two-bus power flow example visualized in *PowerWorld*. Source: U.S.-Canada Power System Outage Task Force, 2004/United States Department of Energy/Public Domain. 350
- Figure 12.13** Two-bus power flow example solved in *PowerWorld*TM. 353

- Figure 12.14** The unrealistic low-voltage solution to the two-bus power flow example. 354
- Figure 12.15** Convergence regions for the two-bus power flow example. Source: Courtesy of Tom Overbye. 354
- Figure 12.16** Three-bus network. 359
- Figure 12.17** Radial distribution branch for the derivation of LinDistFlow equations. 364
- Figure 12.18** Qualitative association of voltage drop $\Delta V = V_1 - V_2$ and angle difference $\Delta\delta = \delta_1 - \delta_2$ with the direction of real and reactive power flow, based on Eqs. (12.5) and (12.8). 366
- Figure 12.19** Three-bus example for Problem 12.4. 367
- Figure 13.1** Toy example illustrating $N-1$ security. 375
- Figure 13.2** Stable and unstable equilibria: (a) bowl and marble in stable equilibrium; (b) ruler in unstable equilibrium. 378
- Figure 13.3** Power transmitted versus power angle. 380
- Figure 13.4** (a) Deep and (b) shallow equilibria. 381
- Figure 13.5** Power generated as a function of power angle. 384
- Figure 13.6** Damped harmonic oscillation in $\delta(t)$, as modeled and measured between two locations in Texas, illustrating generator swing and ringdown after the sudden loss of another generator. Source: Courtesy of Mack Grady. 387
- Figure 13.7** Restoring power and “potential energy” $W(\delta)$, where K.E. = kinetic energy; P.E. = potential energy. 388
- Figure 13.8** Equal area criterion. 390
- Figure 13.9** Real power transfer as a function of voltage phase angle difference across an inductive transmission line. 396
- Figure 13.10** Family of $P-V$ curves. 400
- Figure 13.11** $V-Q$ curve showing bus voltage magnitude versus reactive power injection (negative reactive load) for a simple radial system. 401
- Figure 13.12** Family of $V-Q$ curves for different transmission path impedances, illustrating the problem of voltage security. Source: Graphic from *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, U.S.-Canada Power System Outage Task Force, 2004. 403
- Figure 14.1** Mercury arc valve, configured as (a) a half-wave (three anodes) and (b) full-wave (six anodes) rectifier. Source: Wdwd-Own work/Wikipedia/ CC BY 3.0. 407
- Figure 14.2** Basic schematic of a vacuum triode, with labels for cathode, plate (anode), and grid. 408
- Figure 14.3** p-n Junction and electric field. Source: TheNoise/Wikipedia/CC BY SA.3.0. 411
- Figure 14.4** Inverter waveforms. 416
- Figure 14.5** Pulse-width modulation with 8 cycles (a) and 16 cycles (b). Horizontal gray lines represent the average value of each pulse. 417
- Figure 14.6** Full-bridge inverter circuit. 418
- Figure 14.7** PWM signal. 420
- Figure 14.8** Sample volt-VAR droop curve. 423

- Figure 14.9** Sample volt-watt droop curve. The corner points that define the curve are specified based on local utility requirements. 423
- Figure 15.1** Conceptual illustration for a fossil-fuel power plant. 427
- Figure 15.2** Microgrid concept, with generation (circles), storage (triangle), and loads (rectangles) at various scales. Dashed lines indicate optional connections. 450
- Figure 16.1** Time scales in electric grid operation. 455
- Figure 16.2** Making a voltage measurement on a parallel branch. 463
- Figure 16.3** Potential and current transformers. 464
- Figure 16.4** Voltage divider. 464
- Figure 16.5** Oscillations in local a.c. frequency due to generators swinging against each other, rendered observable with early PMUs. Source: From Faulk & Murphy (1994). 466
- Figure 16.6** Recorded waveforms during a transmission fault. Source: NERC (2017)/with permission of North American Electric Reliability Corporation. 467
- Figure 16.7** California Independent System Operator (CAISO) control room. Source: California Independent System Operator. 469

Preface

This book is a labor of love. It is intended to bridge the gap between standard engineering texts and more popularly accessible descriptions of electric power technology. I discovered this gap as a graduate student struggling to understand power systems, which had always fascinated me, but which I now needed to understand properly in the context of my passion, implementing solar energy. Although I had studied physics as an undergraduate, I found the subject of power systems difficult and intimidating.

The available literature seemed to fall into two categories: easy-to-read, qualitative descriptions of the electric grid for the layperson on the one hand, and highly technical books and papers on the other, written for professionals and electrical engineering majors. The second category had the information I needed, but guarded by a layer of impenetrable phasor diagrams and other symbolism that clearly required some sort of initiation.

I was very fortunate to study with renowned experts at the University of California, Berkeley, including Leon Chua and Felix Wu, who were also generous and gifted teachers and brought me up to speed in deciphering the academic and engineering literature. The formative learning experience was a research project beginning in 1989 at several large nuclear and fossil-fueled steam generation plants, where our team interviewed the staff as part of a study on High-Reliability Organizations. My own subsequent research on power distribution took me into the field with five U.S. utilities and one in Germany. Aside from the many intriguing things we learned about the operating culture in these settings, I discovered how clearly the power plant staff could often explain technical concepts about their working systems. Their language was characteristically plain and direct and was always guided by practical considerations, such as, What does this dial tell you? What happens when you push that button?

In hindsight, the defining moment for inspiring this book occurred in the Pittsburgh control room when I revealed my ignorance about reactive power (ironically, having just boasted about my physics degree, to the operators' benign amusement). They generously supplied me with a copy of the plant operating manual, which turned out to contain the single most lucid and comprehensible explanation of electric generators, including reactive power, I had seen. That manual proved to me that one can write about electric power systems in a way that is accessible to audiences who have not undergone the initiation rites of an academic engineering program, but who nevertheless want to get the real story. I imagined there might be other people much like myself—outside the engineering profession or the power industry but vitally concerned with it—who could benefit from such a practical approach.

After finishing my dissertation in 1995, I decided to give it a try. My goal was to write the book that I would have wanted to read as a student. The guiding principle was to assume a minimum of prior knowledge on the part of the readers while trying to relate as much as possible to their direct

experience and intuition. In essence, the book should explain how the electric grid works and what its relevant technical constraints are. It should convey a sense of the complexity of the system as a whole and point to the tools used by power engineers to navigate that complexity. Though it could not be expected to offer a full professional training in the use of these tools, it should prepare readers for the more specialized literature or advanced coursework. Enough historical context should be provided to view technological innovation both from the standpoint of modern opportunities and the challenges presented by the legacy infrastructure.

From the start, I envisioned two main audiences. The first consists of students and researchers who are learning about electricity and power engineering in an academic setting, and who feel that their understanding would be enhanced by a qualitative, conceptual emphasis to complement the quantitative methods stressed in technical courses. This audience might include students of diverse backgrounds or differing levels of preparation, perhaps transferring into an engineering program from other disciplines. Such students often need to solidify their understanding of basic information that is presumed to be second nature to advanced undergraduates in technical fields. As a supplement to standard engineering texts, the first edition aimed to provide a clear and accessible review of units, definitions, and fundamental physical principles; to explain in words the ideas shown by equations; to contextualize information, highlighting connections among different topics and pointing out their relevance; and to offer a glimpse into the practical world of the electric power industry.

The second major audience consists of professionals working in and around the power industry whose educational background may not be in electrical engineering, but who wish to become more familiar with some of the technical details and the theoretical underpinnings of the system they deal with. This group might include analysts, administrators, and managers coming from fields of business, economics, law, or public policy, as well as individuals with technical or multidisciplinary training in areas other than power engineering.

The project was more ambitious than I had imagined, and the first edition took a decade to write—but it also met with more success than I had dared to hope. After 13 years as a professor of Energy Management & Design at Sonoma State University, I returned to Berkeley in 2012 to teach a power systems course in the Department of Electrical Engineering and Computer Science and to direct electric grid research at the California Institute for Energy and Environment. My group's research primarily focused on the application of novel sensing, analytics, and control strategies to facilitate the grid integration of renewable resources. This work gave me the opportunity to connect with and learn from many experts, especially through the North American Synchrophasor Initiative (NASPI). It also reinforced the sense that a growing audience of scholars and practitioners from diverse fields could use help understanding how the electric grid works.

The two-semester course *Introduction to Electric Power Systems* I developed and taught for 10 years gives an overview of the legacy electric grid and highlights opportunities and challenges associated with incorporating new technologies. The course has served engineering majors as well as students at both the undergraduate and graduate level from other departments with an interest in the electric grid and with enough preparation to handle the quantitative assignments. I used the first edition of this book for my class, but had to supplement it with other resources to cover some of the more technical aspects. Over time, my accumulated lecture notes evolved into this substantially expanded second edition, which aims to serve as a reasonably complete stand-alone text for introductory engineering courses like mine.

To this end, some chapters were amended to cover more technical points and problem-solving tools, some material was reorganized, and seven new chapters were added. Along with more formal treatment and quantitative examples for many topics, new material in this second edition

includes Thévenin and Norton equivalents, symmetrical components, ZIP load models, transmission line analysis, transformer and generator modeling, load frequency control, power transfer limits, power electronics, measurement and data, and a summary chapter on generation and storage resources. Finally, it includes some end-of-chapter problems, with solutions made available separately through the publisher's website.

While meeting the needs of a rigorous engineering text, this second edition should not sacrifice accessibility for general audiences. It aspires to satisfy readers who think in mathematical terms, and those who don't. As a rule, conceptual explanations precede mathematical statements, and the more technical portions can be skipped without leaving the reader disoriented. Sections are extensively cross-referenced and intended to be modular. The idea is that the reader can open up the book to any page that piques their curiosity or answers a nagging question and be assisted in locating background information and relevant context elsewhere in the book. There are many plausible ways to organize the material—or likely some subset of it—for a course. Chapters need not be read or covered in sequence. Instructors should be able to select and tailor material to their purposes.

One major change in recent years is the availability of technical content on the Internet, including Wikipedia and YouTube tutorials. Especially for animated visualizations (say, rotating machines!), some of these resources can be immensely valuable. But aside from reliable quality control, what the Internet does not provide is coherent guidance and context. For autodidactic readers, this book aims to serve as a foundation and structure that will support them in the critical consumption and interpretation of other educational materials about electricity-related topics.

This is a great time for many more students, whether inside or outside traditional academic settings, to develop a passion for understanding electric power systems. The electric grid is central to our society, our economy, and our efforts to address global climate change. While being host to a shifting base of energy resources, including inverter-based generation, the infrastructure itself is in the process of being modernized with digital technology. Nobody yet can answer the question, “What does the grid of the future look like?” We urgently need diverse talents and skills applied to this vital problem area. My hope is that this book can make a helpful contribution.

Bishop, CA
February 2024

Alexandra von Meier

Acknowledgments

Many individuals and organizations have made the writing of this text possible. I will always be deeply grateful to my teachers, most especially Gene Rochlin and Felix Wu, for their mentorship. I am also indebted to the many industry professionals who took the time to show me around power systems in the field and teach me about their work.

The writing of the first edition was supported both directly and indirectly by a University of California President's Postdoctoral Fellowship, the University of California Energy Institute (UCEI), the California Energy Commission, and the California Institute for Energy and Environment (CIEE). Carl Blumstein, Ron Hofmann, and Laurie ten Hope made possible a series of tutorials that developed early material for the book.

Cary Berkley, Thomas Harris, Darcy McCormick, and Steve Shoptaugh helped with graphics and assembling the first edition. Special thanks go to Andrew Busby and Wei Lao for their tireless work in wrangling the second edition manuscript and Michael Sankur for creating numerous, greatly improved new figures.

The second edition benefited immeasurably from the thoughtful and creative contributions of my teaching assistants for EE 137AB at UC Berkeley: Dan Arnold, Mohini Bariya, Kyle Brady, Samantha Coday, Roel Dobbe, Laurel Dunn, Ioannis Konstantakopoulos, Veda Krishnan, Jonathan Lee, Keith Moffat, Aminy Ostfeld, Michael Sankur, Samuel Smith, Jaimie Swartz, Michaelangelo Tabone, Insoon Yang, and Eric Yehl. I also learned from every student who asked a question in class or office hours that made me stop and think.

I am immensely grateful to my colleagues and friends who read, discussed, helped improve, and generously contributed to the manuscript over the years. They include Raquel Blanco, Carl Blumstein, Duncan Callaway, Joe Eto, Alex Farrell, Hannah Friedman, John Galloway, Mack Grady, Chris Greacen, Sean Greenwalt, Dianne Hawk, Nicole Hopper, Merrill Jones, Jonathan Lee, Chris Marnay, Andrew McAllister, Alex McEachern, Keith Moffat, Tom Overbye, Liz Ratnam, Michael Sankur, Steve Shoptaugh, Leigh Tesfatsion, Jim Williams, and all the readers who took the time to send feedback. Finally, I thank my husband Mike Kearney for his loving support.

Of course, I am solely responsible for any errors. I imagine this text will remain a work in progress, and suggestions for improving its accuracy and clarity will be warmly welcomed.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/vonmeier/electricpowersystems2



This website includes:

- Solution Manual
- PowerPoint slides

1

Physics of Electricity

1.1 Basic Quantities

This chapter describes the quantities that are essential to our understanding of electricity: charge, voltage, current, resistance, and electric and magnetic fields. Most students of science and engineering find it very hard to gain an intuitive appreciation of these quantities, since they are not part of the way we normally see and make sense of the world around us. Electrical phenomena have a certain mystique that derives from the difficulty of associating them with our direct experience, but also from the knowledge that they embody a potent, fundamental force of nature.

Electric charge is one of the basic dimensions of physical measurement, along with mass, distance, time, and temperature. All units in physics can be expressed as some combination of these terms. Unlike the others, however, charge is more remote from our sensory perception. While we can easily visualize the size of an object, imagine its weight, or anticipate the duration of a process, it is difficult to conceive of “charge” as a tangible phenomenon.

To be sure, electrical processes are vital to our bodies, from cell metabolism to nervous impulses, but we do not usually conceptualize these in terms of electrical quantities or forces. Our most direct and obvious experience of electricity is to receive an electric shock. Here the presence of charge sends such a strong wave of nervous impulses through our body that it produces a distinct and unique sensation. Other firsthand encounters with electricity include hair that defiantly stands on end, a zap from a door knob, and static cling in the laundry. Yet these experiences hardly translate into the context of electric power, where we can witness the *effects* of electricity, such as a glowing light bulb or a rotating motor, while the essential happenings take place silently and concealed within pieces of metal. For the most part, electricity remains an abstraction to us, and we rely on numerical and geometric representations—aided by liberal analogies from other areas of the physical world—to form concepts and develop an intuition about it.

1.1.1 Charge

It was a major scientific accomplishment to integrate an understanding of electricity with fundamental concepts about the microscopic nature of matter. Observations of static electricity like those mentioned earlier were elegantly explained by Benjamin Franklin in the late 1700s as follows: There exist in nature two types of a property called *charge*, arbitrarily labeled “positive” and “negative.” Opposite charges attract each other, while like (similar) charges repel. When certain materials rub together, one type of charge can be transferred by friction and “charge up” objects that subsequently repel objects of the same kind (hair), or attract objects of a different kind (polyester and cotton, for instance).

Through a host of ingenious experiments,¹ scientists arrived at a model of the atom as being composed of smaller individual particles with opposite charges, held together by their electrical attraction. Specifically, the nucleus of an atom, which constitutes the vast majority of its mass, contains *protons* with a positive charge, and is enshrouded by *electrons* with a negative charge. The nucleus also contains neutrons, which resemble protons, except they have no charge. The electric attraction between protons and electrons just balances the electrons' natural tendency to escape, which results from both their rapid movement, or kinetic energy, and their mutual electric repulsion. (The repulsion among protons in the nucleus is overcome by another type of force called the *strong nuclear interaction*, which only acts over very short distances.)

This model explains both why most materials exhibit no obvious electrical properties, and how they can become “charged” under certain circumstances: The opposite charges carried by electrons and protons are equivalent in magnitude, and when electrons and protons are present in equal numbers (as they are in a normal atom), these charges “cancel” each other in terms of their effect on their environment. Thus, from the outside, the entire atom appears as if it had no charge whatsoever; it is *electrically neutral*.

Yet individual electrons can sometimes escape from their atoms and travel elsewhere. Friction, for instance, can cause electrons to be transferred from one material into another. As a result, the material with excess electrons becomes *negatively charged* and the material with a deficit of electrons becomes *positively charged* (since the positive charge of its protons is no longer compensated). The ability of electrons to travel also explains the phenomenon of *electric current*, as we will see shortly.

Some atoms or groups of atoms (molecules) naturally occur with a net charge because they contain an imbalanced number of protons and electrons; they are called *ions*. The propensity of an atom or molecule to become an ion—namely, to release electrons or accept additional ones—results from peculiarities in the geometric pattern by which electrons occupy the space around the nuclei. Even electrically neutral molecules can have a local appearance of charge that results from imbalances in the spatial distribution of electrons—that is, electrons favoring one side over the other side of the molecule. These electrical phenomena within molecules determine most of the physical and chemical properties of all the substances we know.²

On the microscopic level, one deals with fundamental units of charge (that of a single electron or proton). The practical unit of charge in the context of electric power is the *coulomb* (C). One coulomb corresponds to the charge of 6.25×10^{18} protons. Stated the other way around, one proton has a charge of 1.6×10^{-19} C. One electron has a negative charge of the same magnitude, -1.6×10^{-19} C. In equations, charge is conventionally denoted by the symbol Q or q.

1.1.2 Potential or Voltage

Because like charges repel, charge has a natural tendency to “spread out.” A local accumulation or deficit of electrons causes a certain “discomfort” or “tension.”³ Unless physically restricted, these charges will tend to move in such a way as to relieve the local imbalance. In rigorous physical terms,

¹ Almost any introductory physics text will provide examples. For an explanation of the basic physical concepts of electricity, I recommend Paul Hewitt, *Conceptual Physics*, currently in the 13th edition (Pearson, 2022).

² For example, water owes its amazing liquidity and density at room temperature to the electrical attraction among its neutral molecules that results from each molecule being polarized: casually speaking, the electrons prefer to hang out near the oxygen atom as opposed to the hydrogen atoms of H₂O; a chemist would say that oxygen has a greater *electronegativity* than hydrogen. The resulting attraction between these polarized ends of molecules is called a *hydrogen bond*, which is essential to all aspects of our physical life.

³ The term *tension* is actually synonymous with voltage or potential, mainly in British usage.

the degree of discomfort is expressed as a level of *energy*. This energy (strictly, electrical potential energy), said to be “held” or “possessed” by a charge, is analogous to the mechanical potential energy possessed by a massive object when it is elevated above the ground: we might say that, by virtue of its height, the object has an inherent potential to fall down. A state of lower energy—closer to the ground, or farther away from like charges—represents a more “comfortable” state, with a smaller potential fall.

The potential energy held by an object or charge in a particular location can be specified in two ways that are physically equivalent: first, it is the *work*⁴ that would be required in order to move the object or charge *to* that location (from a reference location or *ground*, see below). For example, it takes work to lift an object; it also takes work to bring an electron near an accumulation of more electrons. Alternatively, the potential energy is the work the object or charge would do in order to move *from* that location back to the reference location, through interacting with the objects in its way. For example, a weight suspended by a rubber band will stretch the rubber band in order to move downward with the pull of gravity (from higher to lower gravitational potential). A charge moving toward a more comfortable location might do work by producing heat in the wire through which it flows.

This notion of work is crucial because, as we will see later, it represents the physical basis of transferring and utilizing electrical energy. In order to make this “work” a useful and unambiguous measure, some proper definitions are necessary. The first is to explicitly distinguish the contributions of charge and potential to the total amount of work or energy transferred. Clearly, the amount of work in either direction (higher or lower potential) depends on the amount of mass or charge involved. For example, a heavy weight would stretch a rubber band farther, or even break it. Similarly, a greater charge will do more work in order to move to a lower potential.

On the other hand, we also wish to characterize the location proper, independent of the object or charge there. Thus, we establish the rigorous definition of the electric *potential*, which is synonymous with *voltage* (but more formal). The electric potential is the potential energy possessed by a charge at the location in question, relative to a reference location, divided by the amount of its charge. Casually speaking, we might say that the potential represents a measure of how comfortable or uncomfortable it would be for any charge to reside at that location. A potential or voltage can be positive or negative. A positive voltage implies that a positive charge would be repelled and a negative charge would be attracted to the location; a negative voltage implies the opposite.

Furthermore, we must be careful to specify the “reference” location: namely, the place where the object or charge was moved from or to. In the mechanical context, we specify the height *above ground level*. In electricity, we refer to an electrically neutral place, real or abstract, with zero or *ground potential*. Theoretically, we can call this a place where no other charges are present to exert any forces, such as deep space. In practice, ground potential is any place where positive and negative charges are balanced and their influences cancel. When describing the potential at a single location, it is implicitly the potential *difference* between this and the neutral location. However, potential can also be specified as a difference between two locations of which neither is neutral, like a difference in height.

Because electric potential or voltage equals energy per charge, the units of voltage are equivalent to units of energy divided by units of charge. These units are *volts* (V). One volt is equivalent to one joule per coulomb, where the joule is a standard unit of work or energy.⁵

⁴ In physics, work is equivalent to and measured in the same units as energy, with the implied sense of exerting a force to “push” or “pull” something over some distance (work = force × distance).

⁵ A joule can be expressed as a watt-second. 1 kilowatt-hour (kWh) = 3.6×10^6 J, since there are 3600 seconds in an hour.

The notion of a difference always remains implicit in the measurement of volts. A statement like “this wire is at a voltage of 100 volts” means “this wire is at a voltage of 100 volts *relative to ground*,” or “the voltage *difference between the wire and the ground* is 100 volts.” By contrast, if we say “the battery has a voltage of 1.5 volts,” we mean that “the voltage *difference between the two terminals* of the battery is 1.5 volts.” Note that the latter statement does not tell us the potential of either terminal in relation to ground, which depends on the type of battery and what else it is connected to.

In equations, voltage is conventionally denoted by E , e , V , or v (in a rare and inelegant instance of using the same letter for both the symbol of the quantity and its unit of measurement).

1.1.3 Ground

The term ground has a very important and specific meaning in the context of electric circuits: it is an electrically neutral place, meaning that it has zero voltage or potential, which moreover has the ability to absorb excesses of either positive or negative charge and disperse them so as to *remain* neutral regardless of what might be electrically connected to it.

The literal ground outdoors has this ability because the Earth as a whole acts as a vast reservoir of charge and is electrically neutral, and because most soils allow electric charge to travel away from any local accumulation, in the form of ions. The term earth is synonymous with ground, especially in British usage. A circuit “ground” is constructed simply by creating a pathway for charge into the earth. In the home, this is often done by attaching a wire to metal water pipes. In power systems, ground wires, capable of carrying large currents if necessary, are specifically dug into the earth.

The quality of a ground connection (quantified in terms of *impedance*, Section 3.3) can be important in practice. It depends on the surface area across which the ground wire can interact with the soil chemistry.

1.1.4 Conductivity

To understand conductivity, we must return to the microscopic view of matter. In most materials, electrons are bound to their atoms or molecules by the attraction to the protons in the nuclei. We have mentioned how special conditions such as friction can cause electrons to escape. In certain materials, some number of electrons are always free to travel. As a result, the material is able to *conduct* electricity. When a charge (i.e., an excess or deficit of electrons) is applied to one side of such a conducting material, the electrons throughout will realign themselves, spreading out by virtue of their mutual repulsion, and thus conduct the charge to the other side.

For this to happen, an individual electron need not travel very far. We can imagine each electron moving a little to the side, giving its neighbor a repulsive “shove,” and this shove propagating through the conducting material like a wave of falling dominoes.

The most important conducting materials in our context are metals. The microscopic structure of metals is such that some electrons are always free to travel throughout a fixed lattice of positive ions (the atomic nuclei surrounded by the remaining, tightly bound electrons).⁶ While all metals conduct, their conductivity varies quantitatively depending on the ease with which electrons can travel, or the extent to which their movement tends to be hampered by microscopic forces and collisions inside the material.

⁶ This property can be understood through the periodic table of the elements, which identifies metals as being those types of atoms with one or a few electrons dwelling alone in more distant locations from the nucleus (orbitals), from where they are easily removed (ionized) so as to become free electrons.

Besides metals, there are other types of material that conduct electricity. One is water, or any other fluid, with dissolved ions (such as salt or minerals). In this case, it is not electrons but entire charged molecules that travel through the fluid to carry a current. Only small concentrations of ions are needed to make water conductive; while pure distilled water does not conduct electricity, normal tap water and rain water conduct all too well.⁷

Some materials, including air, can also become temporarily conductive through ionization. In the presence of a very strong potential gradient (defined as an *electric field* in Section 1.5), or intense heat, some electrons are stripped from their molecules and become free to travel. A gas in this state is called a *plasma*. Plasmas exist inside stars, nuclear fusion reactors, and fluorescent lights. More often, though, ionization is local and transient: it occurs along a distinct trail, since ionized molecules incite their neighbors to do the same, and charge flows along this trail until the potential difference (charge imbalance) is neutralized. This is precisely what happens in an electric spark across an air gap, an arc between power lines, or a lightning bolt.⁸

In many engineering situations, it is important to predict just when ionization might occur; namely, how great a potential difference over how short a distance will cause “arcing.” For air, this varies according to temperature and especially humidity, as well as the presence of other substances like salt suspended in the air. Exact figures for the *ionizing potential* can be found in engineering tables. For units of conductivity and the relationship to resistance, see Section 1.2.

Finally, some materials can become *superconducting*, generally at very low temperatures. Here, electrons undergo a peculiar energetic transition that allows them to travel with extreme ease, unimpeded by any obstructive forces or collisions. Thus, electrons in the superconducting state do no work on anything in their path, and therefore lose no energy. Some ceramic materials attain superconductivity at a temperature easily sustained by cooling with liquid nitrogen (with a boiling point of minus 321°F), which is referred to as “high temperature” in contrast to previously known superconductivity near absolute zero.⁹ While liquid nitrogen is quite cheap in a research setting, large-scale cooling systems for enabling superconductivity in electric power applications are still widely considered too expensive and cumbersome to be of broad practical interest. This may be changing; see Section 7.2.5 on superconducting transmission. Another conceivable application of superconductivity in power systems is superconducting magnetic energy storage (SMES).

1.1.5 Current

When charge travels through a material, an *electric current* is said to flow. The current is quantified in terms of the number of electrons (or equivalent charge, in the case of ions) moving past a given point in the material in a certain period of time. In other words, current is a *flow rate* of charge. In this way, electric current is analogous to a flow rate of water (say, in gallons per minute) or natural gas (cubic feet per second).

These analogies are also helpful in remembering the distinction between current and voltage. Voltage would be analogous to a height difference (say, between a water reservoir and the downhill end of a pipe), or to a pressure difference (between two ends of a gas pipeline). Intuitively, voltage

⁷ In fact, conductivity is used as an indicator of water purity. Of course, it says nothing about the kind of ions present, only the amount.

⁸ The ionization trail is visible because, as the electrons return to their normal state, the balance of their energy is released in the form of light.

⁹ The first high-temperature superconductor to be discovered was yttrium-barium-copper oxide. There are now a variety of such materials, including some (e.g., bismuth-strontium-calcium-copper oxide) that do not use rare-earth elements.

is a measure of “how badly the stuff wants to get there,” and current is a measure of “how much stuff is actually going.”

Current is conventionally denoted by the symbol I or i and is measured in units of *ampère* (A), often called “amps.” Since current represents a flow rate of charge, the units of current are equivalent to units of charge divided by units of time. Thus, one ampere equals one coulomb per second.

A subject that often causes confusion is the “direction” in which current flows, though in practice, having an accurate picture of this is not all that important. Most often, the reasons one is concerned with current have to do with the amount of power transferred or the amount of heating of the wires, neither of which depend on direction.

When in doubt, we can always refer back to the fact that opposite charges attract and like charges repel. Thus, a positive charge will be attracted by a negative potential, and hence flow toward it, and *vice versa*: electrons, which have negative charge, flow toward a positive potential or voltage. In a mathematical sense, negative charge flowing in one direction is equivalent to positive charge flowing the opposite way. Indeed, our practical representation of electric current does not distinguish between these two physical phenomena. For example, the current flowing through a lead–acid battery at various times consists of negative electrons in the terminals and wires, and positive ions in the battery fluid; yet these flows are thought of as the same current.

In circuit analysis, it often becomes necessary to define a direction of current flow, so as to know when to add and when to subtract currents that meet on a section of the circuit. The general convention is to label a current flow as “positive” in the direction from positive toward negative potential (as if a positive charge were flowing). Once this labeling has been chosen, all currents in the circuit will be computed as positive or negative so as to be consistent with that requirement (positive currents will always point toward lower potential). However, the convention is arbitrary in that one can define the currents throughout an entire circuit “backward,” and obtain just as “correct” a result. In other words, for purposes of calculation, the quantity “current” need not indicate the actual physical direction of traveling charge.

In the power systems context, the notion of directionality is more complicated (and less revealing) because the physical direction of current flow actually alternates (see Section 3.1). Instead, to capture the relationship between two currents (whether they add or subtract), the concept of *phase*, or relative timing, is used.

As for the speed at which current propagates, it is often said that current travels at the speed of light (186,000 mi/s). While this is not quite accurate (just as the speed of light actually varies in different materials), it is usually sufficient to know that current travels very fast.

Conceptually, it is important to recognize that what is traveling at such a high speed is the pulse or signal of the current, not individual electrons departing at one end and arriving at the other end of the conductor. Rather, the electrons inside a metal conductor continually move in a more or less random way, wiggling around in different directions at a speed related to the temperature of the material. They then receive a “shove” in one direction by the electric field (see Section 1.5.2). We can imagine this shove propagating by way of the electrical repulsion among electrons: each electron need not travel a long distance, just enough to push its neighbor over a bit, which in turn pushes its neighbor, and so on.¹⁰ This chain reaction creates a more orderly motion of charge, as

¹⁰ This intuitive description is not strictly correct in terms of quantum mechanics, but adequate for the purposes of this book.

opposed to the usual random motion, and is observed macroscopically as the current. It is the signal to “move over” that propagates at essentially the speed of light.¹¹

The question of the propagation speed of electric current only becomes relevant when the distance to be covered is so large that the time it takes for a current pulse to travel from one point to another is significant compared to other timing parameters of the circuit. This can be the case for electric transmission lines that extend over many hundreds of miles.¹² However, we will not deal with this problem explicitly (see Section 12.3.3 for more on how we treat the concept of time in power systems). A circuit that is sufficiently small so that the speed of current is not an issue is called a *lumped circuit*. Circuits are treated as lumped circuits unless otherwise stated.

1.2 Ohm's Law

It is intuitive that voltage and current would be somehow related. For example, if the potential difference between two ends of a wire is increased, we would expect a greater current to flow, just like the flow rate of gas through a pipeline increases when a greater pressure difference is applied. For most materials, including metallic conductors, this relationship between voltage and current is linear: as the potential difference between the two ends of the conductor increases, the current through the conductor increases proportionally. This statement is expressed in Ohm's law,

$$V = IR \quad (1.1)$$

where V is the voltage, I is the current, and R is the proportionality constant called the *resistance*.¹³ Materials for which Ohm's law holds are called *ohmic*.

Note that Ohm's law is written in *scalar* quantities, that is, plain numbers without a sense of directionality, despite the fact that both voltage and current have an associated direction. We may do this because the proportionality between voltage and current holds regardless of direction (for ohmic materials). With the implicit understanding that the directions of voltage and current align, Ohm's law can be stated generally and does not require reference to any particular points in space until we apply it to a specific situation. Note also that there is no mention of time in Ohm's law. On the relevant time scale for standard circuit analysis, we take voltages and currents to exist and their mutual causal effects to manifest *instantaneously*. In other words, we ignore any “propagation” of voltage or current from one point to another.

1.2.1 Resistance

To say that Ohm's law is true for a particular object (conductor, electrical device, or circuit element) is to say that the resistance of this object does not vary with respect to current and voltage. In the context of analyzing an electric circuit, the resistance is assumed to be constant. Most devices in this book contain metal conductors and obey Ohm's law. They are called *linear* circuit elements.

¹¹ We can draw an analogy with an ocean wave: the water itself moves essentially up and down, and it is the “signal” to move up and down that propagates across the surface, at a speed much faster than the bulk motion of water.

¹² For example, traveling down a 500-mi transmission line at the speed of light takes 2.7 ms. Compared to the rate at which alternating current changes direction (60 times per second, or every 16.7 ms), this corresponds to one-sixth of a cycle, which is not negligible.

¹³ As we will appreciate in Chapter 3, Eq. (1.1) describes the special case of *direct current*, where all quantities are real numbers. For the more general case where current and voltage alternate, Ohm's law is written with complex numbers as $V = IZ$, where Z is the complex *impedance* (Section 3.3).

Certain materials and electronic devices exhibit a nonlinear relationship between current and voltage, that is, their effective resistance varies depending on the voltage applied. These nonlinear devices have specialized applications and will not be discussed in this chapter, but we encounter them in Chapter 14.

The resistance of an object that obeys Ohm's law may still change depending on external factors, but this is usually a slow change that can be neglected in the context of describing an electric circuit at any given moment. Most importantly, resistance does tend to vary with temperature. For example, the resistance of a copper wire increases as it heats up. However, the copper wire still obeys Ohm's law at any given temperature.¹⁴ In most operating regimes, these variations are negligible. In any situation where changes in resistance are significant, this should be explicitly mentioned. Thus, whenever one encounters the term "resistance" without further elaboration, it is safe to assume that within the given context, this resistance is a fixed, unchanging property of the object in question.

Resistance depends on an object's material composition as well as its size and shape. For a wire, resistance increases with length, and decreases with cross-sectional area. Again, the analogy to a gas or water pipe is handy: we know that a pipe will allow a higher flow rate for the same pressure difference if it has a greater diameter, while the flow rate will decrease with the length of the pipe. This is due to *friction* in the pipe, and in fact, an analogous "friction" occurs when an electric current travels through a material.

The friction can be explained by referring to the microscopic movement of electrons or ions, and noting that they interact or collide with other particles in the material as they go. The resulting forces tend to impede the movement of the charge carriers and in effect limit the rate at which they pass. These forces vary for different materials because of the different spatial arrangements of electrons and nuclei, and they determine the material's ability to conduct.

This *intrinsic* material property (independent of size or shape) is called *resistivity* and is denoted by ρ (Greek lowercase rho). The actual resistance of an object is given by the resistivity multiplied by the length of the object (l) and divided by its cross-sectional area (A):^{15,16}

$$R = \frac{\rho l}{A} \quad (1.2)$$

The units of resistance are *ohms*, abbreviated Ω (Greek capital omega). By rearranging Ohm's law in Eq. (1.1), we see that resistance equals voltage divided by current ($R = V/I$). Units of resistance are thus equivalent to units of voltage divided by units of current. By definition, one ohm equals one volt per ampère ($1 \Omega = 1 \text{ V}/1 \text{ A}$).¹⁷

Example

A light bulb with resistance $R = 6 \Omega$ is connected to a 3-V battery. From Ohm's law, the current is $I = V/R = 3 \text{ V}/6 \Omega = 0.5 \text{ A}$.

The units of resistivity are ohm-meters ($\Omega\text{-m}$), which can be reconstructed through the preceding formula: when ohm-meters are multiplied by meters (for l) and divided by square meters (for area A), the result is simply ohms.

¹⁴ If we graph V versus I , Ohm's law requires that the graph be a straight line. With temperature, the slope of this line may change.

¹⁵ A is unfortunately the same letter used to abbreviate ampère; we italicize it here when used as a symbol for area.

¹⁶ For an irregularly shaped object, one would compute a three-dimensional integral, but luckily power systems are full of cylindrical shapes.

¹⁷ Note again that the letter V is performing an awkward double duty as a symbol for a quantity and its unit of measurement, although it is often italicized when used as a symbol.

Resistivity, which is an intrinsic property of a material, is not to be confused with the *resistance per unit length* (usually of a wire), quoted in units of ohms *per* meter (Ω/m). The latter measure already takes into account the wire diameter; it represents, in effect, the quantity ρ/A . The resistivities of different materials in $\Omega\text{-m}$ can be found in physical reference tables.

Engineering tables that list the resistance of particular transmission line conductors often distinguish a.c. versus d.c. resistance values in separate columns, and specify the a.c. frequency. The reason for this specification is the *skin effect*, by which alternating current at higher frequencies tends to flow primarily near the surface or “skin” of the conductor. This phenomenon is caused by the interaction of the alternating current with the changing magnetic field internal to the conductor (see also Section 9.1.1), which is stronger near the center. The skin effect reduces the effective cross-sectional area available to the current, and thus increases the effective resistance slightly. It is more significant for conductors of larger diameter.

1.2.2 Conductance

It is sometimes convenient to refer to the resistive property of a material or object in the inverse, as *conductivity* or *conductance*. Conductivity is the inverse of resistivity and is denoted by σ (Greek lowercase sigma): $\sigma = 1/\rho$. For the case of a simple resistor, conductance is the reciprocal of resistance and is usually denoted by G (sometimes g), where $G = 1/R$.¹⁸ The standard unit of conductance is the siemens (S), where $1\text{ S} = 1/\Omega$. Not without humor, this unit has sometimes been called *mho*. Conductance is related to conductivity by

$$G = \frac{\sigma A}{l} \quad (1.3)$$

and the units of σ are thus S/m .

For the special case of an *insulator*, the conductance is zero and the resistance is infinite. For the special case of a superconductor, the resistance is zero and the conductance is, theoretically, infinite. In practice, while the amount of current in a superconductor can be extremely large, it is eventually constrained by the number of electrons available and thus not exactly infinite, but it is unconstrained by resistance.

Example

Consider two power extension cords, one with twice the wire diameter of the other. If the cords are of the same length and same material, how do their resistances compare?

Since resistance is inversely proportional to area, the smaller wire will have four times the resistance. We can see this through the formula $R = \rho l/A$, where ρ and l are the same for both. Thus, using the subscripts 1 and 2 to refer to the two cords, we can write $R_1/R_2 = A_2/A_1$. The areas are given by the familiar geometry formula, $A = \pi(d/2)^2$ (where $\pi = 3.1415 \dots$), which includes the square of the diameter or radius. If the length of either cord were doubled, its resistance would also double.

To put some numbers to this example, consider a typical 25-ft, 16-gauge extension cord, made of a copper conductor. The cross-sectional area of 16-gauge wire is 1.31 mm^2 (or $1.31 \times 10^{-8} \text{ m}^2$) and the resistivity of copper is $\rho = 1.76 \times 10^{-8} \Omega\text{-m}$. The resistance per unit length of 16-gauge copper wire is $0.0134 \Omega/\text{m}$, and a 25-ft length of this single wire has a resistance of 0.102Ω . By contrast, a 10-gauge copper wire of the same length, which has about twice the diameter, has a resistance of only 0.025Ω .

¹⁸ See Section 3.3 for the complex case that includes both resistance and reactance.

Because an extension cord comprises two conductors—one for the forward and one for the return current—the effective length and resistance are twice the above value for each case.

1.2.3 Insulation

Insulating materials are used in electric devices to keep current from flowing where it is not desired. They are simply materials with a sufficiently high resistance (or sufficiently low conductance), also known as *dielectric* materials. Typically, plastics or ceramics are used. When an insulator is functional, its resistance is infinite, or conductance is zero, so that zero current flows through it.

Any insulator has a specific voltage regime within which it can be expected to perform. If the voltage difference between two sides of the insulator becomes too large, its insulating properties may break down due to microscopic changes in the material, where it actually becomes conducting. Generally, the thicker the insulator, the higher the voltage difference it can sustain. However, temperature can also be important; for example, plastic wire insulation may melt if the wire becomes too hot.

The insulators often seen on high-voltage equipment consist of strings of ceramic bells, holding the energized wires away from other components (e.g., transmission towers or transformers). The shape of these bells serves to inhibit the formation of arcs or creeping currents along their surface. The number of bells is roughly proportional to the voltage level (insulating to roughly 10 kV per bell), though it also depends on climate. For example, the presence of salt water droplets in coastal air encourages ionization and therefore requires more insulation to prevent faults.

1.3 Circuit Fundamentals

1.3.1 Static Charge

A current can only flow as long as a potential difference is sustained; in other words, the flowing charge must be replenished. Therefore, some currents have a very short duration. For example, a lightning bolt lasts only a fraction of a second, until the charge imbalance between the clouds and the ground is neutralized.

When charge accumulates in one place, it is called *static charge* because it is not moving. The reason charge remains static is that it lacks a conducting pathway that enables it to flow toward its opposite charge. When we receive a shock from static electricity—for example, by touching a doorknob—our body is providing just such a pathway. In this example, our body is charged through friction, often on a synthetic carpet, and this charge returns to the ground via the door-knob (the carpet only gives off electrons by rubbing, but does not allow them to flow back). As our fingers approach the doorknob, the air in between is actually ionized momentarily, producing a tiny arc that causes the painful sensation.¹⁹ Static electricity occurs mostly in dry weather, since moisture on the surface of objects makes them sufficiently conductive to prevent accumulations of charge.

¹⁹ Charge will accumulate more densely in the point, being attracted to the opposite charge across the gap. The charge density in turn affects the gradient of the electric potential across the gap, which is what causes the ionization. Therefore, approaching the doorknob with a flat hand can prevent the formation of an arc, and charge will simply flow (unnoticeably) after the contact has been made. This is also why lightning arresters work: a particularly pointed object like a metal rod will “attract” an electric arc toward its high charge density. By the same token, lightning tends to strike tall trees and transmission towers.

However startling and uncomfortable, static electricity encountered in everyday situations is harmless because the amount of charge available is so small,²⁰ and it is not being replenished. This is true despite the fact that very high voltages can be involved (recall that voltage is energy *per* charge), but these voltages drop instantaneously as soon as the contact is made.

1.3.2 Closing a Circuit

In order to produce a sustained flow of current, the potential difference must be maintained. This is achieved by providing a pathway to “recycle” charge to its origin, and a mechanism (called an *electromotive force*, or *emf*²¹) that compels the charge to return to the less “comfortable” potential. Such a setup constitutes an electric circuit.

A simple example is a battery connected with two wires to a light bulb. The chemical forces inside the battery do work on the charge to move it to the terminals, where an electric potential is produced and sustained. Specifically, electrons are moved to the negative terminal, and positive ions are moved to the positive electrode, where they produce a deficit of electrons in the positive battery terminal. The wires then provide a path for electrons to flow from the negative to the positive terminal. Because the positive potential is so attractive, these electrons even do work by flowing through the resistive light bulb, causing it to heat up and glow. As soon as the electrons arrive at the positive terminal, they are “lifted” again to the negative potential, allowing the current to continue flowing. In analogy with flowing water, the wires are like pipes that carry water downhill and the battery is like a pump that returns the water to the uphill end of the circuit.

When the wires are connected to form a complete loop, they make a *closed circuit*. If the wire were cut, this would create an *open circuit*, and the current would cease to flow. In practice, circuits are opened and closed by means of switches that make and break electrical contacts.

1.3.3 Voltage Drop

In describing circuits, it is often desirable to specify the voltage at particular points along the way. The difference in voltage between two points in a circuit is referred to as the *voltage drop* across the wire or other component in between. As in Ohm’s law, $V = IR$, this voltage drop is proportional to the current flowing through the component, multiplied by its resistance. The term “voltage drop” is synonymous with “voltage difference,” except that it implies this voltage difference is an unintended result of current flowing through a wire or other electrical device.

As in the analogy of water pipes running downhill, the voltage drops continuously throughout a circuit, from one terminal of the emf to the other. However, just like the slope of the pipes may change, the voltage does not necessarily drop at a steady rate. Rather, depending on the resistance of a given circuit component, the voltage drop across it will be more or less: a component with high resistance will sustain a greater voltage drop, whereas a component with low resistance such as a conducting wire will have a smaller voltage drop across it, perhaps so small as to be negligible in a given context. For small circuits, it is often reasonable to assume that the wire’s resistance is zero, and that therefore the voltage is the same all the way along the wire. In power systems, however, where transmission and distribution lines cover long distances, the voltage drop across them is significant and indeed accounts for some important aspects of how these systems function.

20 The same is not true of electrical equipment that has been specifically designed to hold a very large amount of static charge!

21 Unrelated to the EMF that stands for “electromagnetic fields.”

Importantly, electrical current (or flow rate of charge) also affects the voltage drop. This is a big departure from the water analogy, where the flow rate of water through a pipe has no effect on the elevation difference. For example, at times of high electric demand and thus high current flow, the voltage drop along transmission and distribution lines is greater; that is, the voltage drops more rapidly with distance. If this condition cannot be compensated for by other adjustments in the system (see Section 7.4), customers experience lower voltage levels associated with dimmer lights and impaired equipment performance, known as “brownouts.” Similarly, if a piece of heavy power equipment is connected through a long extension cord with too high a resistance, the voltage drop along this cord can result in damage to the motor from excessively low voltage at the far end.

Example

Suppose the current in the 25-ft, 16-gauge cord from the previous example is 5 A. What is the voltage drop along each conductor?

The voltage drop V_d in the wire is given by Ohm’s law, $V_d = IR$. Using $R = 0.102\Omega$, $V_d = 5\text{ A} \cdot 0.102\Omega = 0.51\text{ V}$ for each 25-foot length of wire.

An appliance plugged into the end of this extension cord would see the voltage difference supplied by the wall outlet diminished by this voltage drop twice, once for each of the two conductors in the cord.

Thus, if the wall outlet supplies 120 V, this would be drop to about 119 V as seen by the appliance at the end of the cord when a current of 5 A is flowing.

This very simple example, where the extension cord is characterized by a resistance R , allows us to use alternating-current quantities without special definitions or caveats.

1.3.4 Electric Shock

Any situation where a high voltage is sustained by an electromotive force (or a very large accumulation of charge) constitutes a shock hazard. Our bodies are not noticeably affected by being “charged up” or raised to a potential above ground, just as birds can sit on a single power line. Rather, harm is done when a current flows through our body. A current as small as a few milliamperes across the human heart can be lethal.²² For current to flow through an object, there must be a voltage drop across it. In other words, our body must be simultaneously in contact with two sources of different potential—for example, a power line and the ground.

Though it is the current that causes biological damage, Ohm’s law indicates that shock hazard is roughly proportional to the voltage encountered. However, the resistance is also important. On an electrical path through the human body, the greatest resistance is on the surface of the skin and clothing, while our interior conducts very well. Thus, the severity of a shock received from a particular voltage can vary, depending on how sweaty one’s palms are, or what type of shoes one is wearing.

The physical principles of electric current can be applied to suggest a number of practical precautions for reducing electric-shock hazards. For example, when touching an object at a single high voltage, we are safe as long as we are insulated from the ground. A wooden ladder might serve this purpose at home, while utility linemen often work on “hot” equipment out of raised plastic “buckets.” Linemen can also insulate themselves from the high-voltage source by wearing

²² A saying goes, “It’s the volts that jolts, and the mils that kills.”

special rubber gloves, which are commonly used for work on up to 12 kV. The important thing is to know the capability of the insulator in relation to the voltage encountered.

A different safety measure often used by electricians when touching a questionable component (such as a wire that might be energized) is to make contact with ground potential with the same hand, for example, by touching the little finger to the wall. In this way, a path of low resistance is created through the hand, which will greatly reduce the current flowing through the rest of the body and especially across the heart. Though the hand might be injured (improbable at household voltage), such a shock is far less likely to be lethal.

Around high-voltage equipment, in order to avoid the possibility of touching two objects at different potentials with a current pathway across the heart, a common practice is to “keep one hand in your pocket.” Near very high potentials, where the concern is not just about touching equipment, but even drawing an arc across the air, the advice is to “keep both hands in your pockets” so as to avoid creating a point with high charge density to attract an arc.

Finally, another factor to consider is the muscular contraction that often occurs in response to an electric shock. Thus, a potentially energized wire is better touched with the back of the hand, so as to prevent involuntary closing of the hand around it. If a person is in contact with an energized source, similar precautions should be exercised in removing them, lest there be additional casualties. If available, a device like a wooden stick would be ideal; in the worst case, kicking is preferable to grabbing.

1.4 Resistive Heating

Whenever an electric current flows through a material that has some resistance (i.e., anything but a superconductor), it creates heat. This *resistive heating* is the result of “friction,” as created by microscopic phenomena such as retarding forces and collisions involving the charge carriers (usually electrons); in formal terminology, the heat corresponds to the work done by the charge carriers in order to travel to a lower potential. This heat generation may be intended by design, as in any heating appliance (e.g., a toaster, an electric space heater, or an electric blanket). Such an appliance essentially consists of a conductor whose resistance is chosen so as to produce the desired amount of resistive heating. In other cases, resistive heating may be undesirable. Power lines are a classic example. For one, their purpose is to transmit energy, not to dissipate it; the energy converted to heat along the way is, in effect, lost (thus the term *resistive losses*). Furthermore, resistive heating of transmission and distribution lines is undesirable, since it causes thermal expansion of the conductors, making them sag. In extreme cases such as fault conditions, resistive heating can literally melt the wires.

1.4.1 Calculating Resistive Heating

There are two simple formulas for calculating the amount of heat dissipated in a resistor (i.e., any object with some resistance). This heat is measured in terms of power, which corresponds to energy per unit time. Thus, we are calculating a *rate* at which energy is being converted into heat inside a conductor. The first formula is

$$P = IV \quad (1.4)$$

where P is the power, I is the current through the resistor, and V is the voltage drop across the resistor.

Power is measured in units of *watts* (W), which correspond to amperes · volts. Thus, a current of 1 A flowing through a resistor across a voltage drop of 1 V produces 1 W of heat. Units of watts can also be expressed as joules per second. To conceptualize the magnitude of a watt, it helps to consider the heat created by a 100-W light bulb, or a 1000-W space heater.

The relationship $P = IV$ makes sense if we recall that voltage is a measure of energy per unit charge, while the current is the flow rate of charge. The product of current and voltage therefore tells us how many electrons are “passing through,” multiplied by the amount of energy each electron loses in the form of heat as it goes, giving an overall rate of heat production. We can write this as

$$\frac{\text{charge}}{\text{time}} \cdot \frac{\text{energy}}{\text{charge}} = \frac{\text{energy}}{\text{time}}$$

and see that, with the charge canceling out, units of current multiplied by units of voltage indeed give us units of power.

The second formula for calculating resistive heating is

$$P = I^2 R \quad (1.5)$$

where P is the power, I is the current, and R is the resistance. This can be derived from Eq. (1.4) by substituting $I \cdot R$ for V according to Ohm’s law. Eq. (1.5) is much more frequently used to calculate resistive heating because the voltage drop across an object is often not known in practice. However, Eq. (1.4) formula has other, more general applications.

As we might infer from the equation, the units of watts also correspond to amperes squared times ohms ($1 \text{ W} = 1 \text{ A}^2 \cdot 1 \Omega$). Thus, a current of 1 A flowing through a wire with 1Ω resistance would heat this wire at a rate of 1 W. Because the current is squared in the equation, 2 A through the same wire would heat it at a rate of 4 W, and so on.

It is absolutely vital to distinguish power from energy. Energy describes the cumulative work done when power is exerted for some duration of time. The most familiar unit of electrical energy is the kilowatt-hour (kWh), representing 1000 W exerted for one hour.

Example

Suppose the 6Ω light bulb from the earlier example, connected to a 3-V battery and drawing a current of 0.5 A, is an old-fashioned incandescent lamp that consists of a thin, resistive tungsten filament inside an evacuated glass bulb.²³ As the filament heats up and glows, it radiates 1.5 W of power. Some of that radiation occurs in the visible spectrum, but the majority is infrared and just produces unwanted heat.

If the light bulb were a modern light-emitting diode (LED), it would not be well characterized by a simple resistance R . However, given R for an electrical device, we may use it to calculate resistive heating without knowing anything else about the device.

Example

A toaster oven draws a current of 6 A at a voltage of 120 V. How much power does it dissipate?

We can find the heat dissipation in two ways. Assuming that the given voltage describes the voltage drop across the device, we may use $P = IV$, which gives $120 \text{ V} \cdot 6 \text{ A} = 720 \text{ W}$.

Alternatively, we could use the resistance, which is 20Ω . In this case, we would write $P = I^2 R$ to find the same answer: $(6\text{A})^2 \cdot 20 \Omega = 720 \text{ W}$.

²³ The vacuum keeps the tungsten from oxidizing. Tungsten is a marvelous metal for this purpose due to its extremely high melting point.

Suppose the toaster oven is operating at 720 W for two hours. How much energy does it use?

Power is energy per unit time. 720 W can also be thought of as “720 watt-hours per hour.” Over the course of two hours, the toaster oven dissipates $720 \text{ W} \cdot 2 \text{ h} = 1440 \text{ Wh} = 1.44 \text{ kWh}$.

The dependence of power on resistance, current, and voltage can be confusing, because it is not always obvious which of these quantities vary and which are fixed in a given situation. Clearly, the power dissipated should increase with increasing voltage and with increasing current. From the formula $P = I^2R$, we might also expect power to increase with increasing resistance, assuming that the current remains constant. However, it may be incorrect to assume that we can vary resistance without also varying the current.

In most situations it is the voltage that remains (approximately) constant. For example, the voltage at a customer’s wall outlet ideally remains at 120 V, regardless of how much power is consumed.²⁴ The resistance is determined by the physical properties of the appliance: its intrinsic design, and, if applicable, a power setting (such as “high” or “low”). Given the standard voltage, then, the resistance determines the amount of current “drawn” by the appliance according to Ohm’s law: higher resistance means lower current, and *vice versa*. In fact, resistance and current are inversely proportional in this case: if one doubles, the other is halved.

What, then, is the effect of resistance on power consumption? The key here is that resistive heating depends on the square of the current, meaning that the power is more sensitive to changes in current than resistance. Therefore, at constant voltage, the effect of a change in current outweighs the effect of the corresponding change in resistance. For example, decreasing the resistance (which, in and of itself, would tend to decrease resistive heating) causes the current to increase, which increases resistive heating by a greater factor. Thus, at constant voltage, the net effect of decreasing resistance is to increase power consumption. An appliance that draws more power has a lower internal resistance.

For an intuitive example, consider the extreme case of a *short circuit*, caused by an effectively zero resistance (usually unintentional). Suppose a thick metal bar were placed across the terminals of a car battery. A very large current would flow, the metal would become very hot, and the battery would be drawn down very rapidly. If a similar experiment were performed on a wall outlet by sticking, say, a fork into it, the high current would hopefully be interrupted by the circuit breaker before either the fork or the wires melted (DO NOT actually try this!). The other extreme case is simply an open circuit, where the two terminals are separate and the resistance of the air between them is infinite: here the current and the power consumption are obviously zero.

Example

Consider two incandescent light bulbs, with resistances of 240 and 480 Ω, respectively. How much power do they each draw when connected to the same 120-V outlet?

First we must compute the current through each bulb, using Ohm’s law. With $V = 120 \text{ V}$ and $R_1 = 240 \Omega$ in $V = IR$, we obtain $I_1 = 0.5 \text{ A}$. For $R_2 = 480 \Omega$, we get $I_2 = 0.25 \text{ A}$.

Now we can use these values for I and R in the power formula, $P = I^2R$, which yields $P_1 = (0.5 \text{ A})^2 \cdot 240 \Omega = 60 \text{ W}$ and $P_2 = (0.25 \text{ A})^2 \cdot 480 \Omega = 30 \text{ W}$.

²⁴ This is generally true because (a) changes in power consumption from an individual appliance are small compared to the total power supplied to the area by the utility and (b) the utility takes active steps to regulate the voltage (see Section 8.6). Dramatic changes in demand do cause changes in voltage, but for the present discussion, it is more instructive to ignore these phenomena and treat voltage as a fixed quantity.

We see that at constant voltage, the bulb with twice the resistance draws half the current and produces half the power.

There are other situations, however, where the current rather than the voltage is constant. Transmission and distribution lines are an important case. Here, resistive heating is (more intuitively) directly proportional to resistance. The crucial difference between the power line and the appliance situation is that for power lines, the current is essentially unaffected by the resistance of the line itself, being determined instead by the load at the end of the line.²⁵ However, the voltage drop along the line (i.e., the difference in voltage between its endpoints, not to be confused with the line voltage relative to ground) is unconstrained and varies depending on current and the line's resistance. Thus, Ohm's law still holds, but it is now I that is fixed. Applying $P = I^2R$ for resistive heating with the current held constant, we see that doubling the resistance of the power line will double resistive losses. Since in practice it is desirable to minimize resistive losses on power transmission and distribution lines, these conductors are chosen with the minimal resistance that is practically and economically feasible.

1.4.2 Transmission Voltage and Resistive Losses

Resistive losses are the reason why increasingly high voltage levels are chosen for power transmission lines. The relationship $P = IV$ tells us that the amount of power transmitted by a line is given by the product of the current flowing through it and its voltage level, as measured either with respect to ground or between two lines or phases of one circuit (i.e., the voltage difference that will be seen by loads connected at the end). We now assume that a certain quantity of power is demanded for the end use, and that appliances will be designed so as to draw whatever current they need, at the given voltage, in order to do their job. There is now a design choice as to what combination of I and V will constitute this power. A higher voltage level implies that in order to transmit the same amount of power, less current needs to flow. Since resistive heating is related to the square of the current, it is highly beneficial from the standpoint of line losses to reduce the current by increasing the voltage.

Before power transformers were available, voltages could not easily be changed throughout the power system. Transmission voltages were therefore limited to levels considered safe for customers. The high currents required caused so much resistive heating that it posed a significant constraint to the expansion of power transmission. When trying to increase power carried at a given voltage, an increasing fraction of the total power is lost on the lines, making transmission uneconomical beyond some point. The increase in losses can be counteracted by reducing the resistance of the conductors, but only at the expense of making them thicker and heavier. Over a century ago, Thomas Edison found the practical limit for transmitting electricity at the level of a few hundred volts to be only a few miles.

With the help of transformers that allow essentially arbitrary voltage conversion (see Chapter 8), transmission voltage levels have grown steadily in conjunction with the geographic expansion of electric power systems, up to about 1000 kilovolts (kV), and with the most common voltages around 100–500 kV. The main factor offsetting the economic benefits of very high voltage is the increased cost and engineering challenge of safe and effective insulation—which applies not just to a transmission line, but all the equipment connected to it.

²⁵ This is because the resistance of the power line itself ought to be small and insignificant compared to that of the appliances at the end, so that a change in the resistance of the line will have a negligible effect on the overall resistance in the circuit, and thus the current flowing through it. The transmission line and the load are *in series*; see Section 2.2.

1.5 Electric and Magnetic Fields

1.5.1 The Field as a Concept

The notion of a *field* is an abstraction initially developed in physics to explain how tangible objects exert forces on each other at a distance, by invisible means. Articulating and quantifying a “field” particularly helps to analyze situations where an object experiences forces of various strengths and directions, depending on its location. Rather than referring to other objects associated with “causing” such forces, it is usually more convenient to just map their hypothetical effects across space. Such a map is then considered to describe properties of the space, even in the absence of an actual object placed within it to experience the results, and this map represents the field.

For example, consider gravity. We know that our body is experiencing a force downward because of the gravitational attraction between it and the Earth. This gravitational force depends on the respective masses of our bodies and the Earth, but it also depends on our location: astronauts traveling into space feel less and less of a pull toward the Earth as they get farther away. Indeed, though the effect is small, we are even slightly “lighter” on a tall mountain or in an airplane at high altitude. If we were interested in extremely accurate measurements of gravity (e.g., to calculate the exact flight path of a ballistic missile), we could construct a map of a “gravitational field” encompassing the entire atmosphere, which would indicate the strength of gravity at any point. This field is caused by the Earth, but does not explicitly refer to the Earth as a mass; rather, it represents in abstract terms the effect of the Earth’s presence. The field also does not refer to any object (such as an astronaut) that it may influence, though such an object’s mass would need to be taken into account in order to calculate the actual force on it. Thus, the gravitational field is a way of mapping the influence of the Earth’s gravity throughout a region of space.

An alternative interpretation is to consider the field as a physical entity in its own right, even though it has no substance of its own. Here we would call gravity a property of the *space itself*, rather than a map telling us about objects such as the Earth in space. Indeed, the field itself can be considered a “thing” rather than a map because it represents potential energy distributed over space. We know of the presence of this potential energy because it does physical *work* on objects: for example, a massive object within the field is accelerated, and in that moment, the energy becomes observable. With this in mind, we can understand the field as the answer to the question: Where does the potential energy reside while we are not observing it?

This notion of the field as a physical entity is a fairly recent one. Whereas classical physics relied on the notion of action-at-a-distance, in which only tangible objects figured as “actors,” the study of very large and very small things in the 20th century has forced us to give up referring to entities that we can touch or readily visualize when talking about how the world works. Instead, modern physics has cultivated more ambiguity and caution in declaring the “reality” of physical phenomena, recognizing that what is accessible to our human perception is perhaps not a definitive standard for what “exists.” Even what once seemed like the most absolute, immutable entities—mass, distance, and time—were proved ultimately changeable and intractable to our intuition by relativity theory and quantum mechanics.

Based on these insights, we might conclude that any quantities we choose to define and measure are in some sense arbitrary patterns superimposed on the vast web of energy and movement that constitutes reality, for the purpose of helping us apprehend this reality with our thoughts. In this sense, we are no more justified in considering a planet a “thing that really exists” than we are a gravitational field. What we really care about as scientists, though, is how useful such a conceptual pattern might be for describing the world in concise terms and making predictions about how

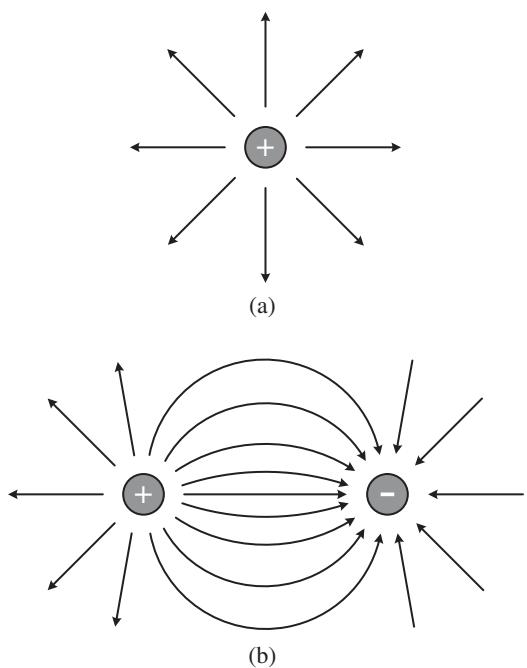
things will behave. By this standard, the notion of a “field” does wonders. Physicists and engineers are therefore accustomed to regarding fields, however devoid of substance, as real, manipulable, and legitimate physical entities just like tangible objects. In any case, the reader should rest assured that it is quite alright to simply accept the “field” as a strange instrument of analysis that grows more palatable with familiarity.

1.5.2 Electric Fields

In Section 1.1.2, we characterized the electric potential as a property of the location at which a charge might find itself. A map of the electric potential would indicate how much potential energy would be possessed by a charge located at any given point. The *electric field* is a similar map, but rather of the electric *force* (such as attraction or repulsion) that would be experienced by that charge at any location. This force is the result of potential differences between locations: the more dramatically the potential varies from one point to the next, the greater the force would be on an electric charge in between these points. In formal terms, the electric field represents the *potential gradient*. The units of electric field strength are volts per meter (V/m).

Consider the electric field created by a single positive charge, just sitting in space. Another positive charge in its vicinity would experience a repulsive force. This repulsive force would increase as the two charges were positioned closer together, or decrease as they moved farther apart; specifically, the electric force drops off at a rate proportional to the square of the distance. This situation can be represented graphically by drawing straight arrows radially outward from the first charge, as in Figure 1.1. Such arrows are referred to as field lines. Their direction indicates the direction that a “test charge,” such as the hypothetical second charge that was introduced, would be pushed or pulled (in this case, straight away). The strength of the force is indicated by the proximity of field lines: the force is stronger where the lines are closer together.

Figure 1.1 Electric field of (a) a single charge and (b) two opposite charge.



This field also indicates what would happen to a negative charge: At any point, it would experience a force of equal strength (assuming equal magnitude of charge), but opposite direction as the positive test charge, since it would be attracted rather than repelled. Thus, a negative test charge would also move along the field lines, only backward. By convention, the direction of the electric field lines is drawn so as to represent the movement of a *positive* test charge.

For a slightly more complex situation, consider the electric field created by a positive and a negative charge, sitting at a fixed distance from each other. We can map the field conceptually by asking, for any location, “What force would be acting on a (positive) test charge if it were placed here?” Each time, the net force on the test charge would be a combination of one attractive force and one repulsive force, in different directions and at different strengths depending on the distance from the respective fixed charges. Graphically, we can construct an image of the field by drawing an arrow in the direction that the charge would be pulled. The arrows for points along the charge’s hypothetical path then combine into continuous field lines. Again, these field lines will be spaced more closely where the force is stronger. This exercise generates the picture in Figure 1.1b.

1.5.3 Magnetic Fields

The pattern of the electric field in Figure 1.1 may be reminiscent to some readers of the pattern that many of us produced once upon a time in science class by sprinkling iron filings on a sheet of paper over a bar magnet. The two phenomena, electric and magnetic forces, are indeed closely linked manifestations of a common underlying physics.

As we know from direct tactile experience, magnets exert force on each other: opposite poles attract, and like poles repel. This is somewhat analogous to the fact that opposite electric charges attract and like charges repel. But, unlike a positive or negative electric charge, a magnetic pole cannot travel individually. There is no such thing as an individual north or south pole (a “monopole” in scientific terms, which has never been found). Every magnet has a north and a south pole. Thus, unlike electric field lines that indicate the direction of movement of an individual test charge, magnetic field lines indicate the *orientation* of a test magnet. The iron filings in the familiar experiment—which become little test magnets since they are magnetized in the presence of the bar magnet—do not move toward one pole or the other, but rotate and align themselves with the direction of the field lines.

It is important to emphasize that, despite the similar shape of field lines, magnetic poles are not analogous to single electric charges sitting in space. Rather than thinking of magnetism as existing in the form of “stuff” like electric charge (which could conceivably be decomposed into its “north” and “south” constituents), it is more appropriate to think of magnetism as an expression of *directionality*, where north is meaningless without south. If you cut a magnet in half, you get two smaller magnets that still each have a north and a south pole.

If we pursued such a division of magnets again and again, down to the level of the smallest particles, we would find that even individual electrons or protons appear as tiny magnets. In ordinary materials, the orientation of all these microscopic magnets varies randomly throughout space, and they therefore do not produce observable magnetic properties at the macroscopic level. It is only in magnetized materials that the direction of these myriad tiny magnets becomes aligned, allowing their magnetic fields to combine to become externally noticeable. This alignment stems from the force magnets exert on each other, and their resulting tendency to position themselves with their north poles all pointing in the same direction. Some substances like magnetite occur naturally with a permanent alignment, making the familiar magnets that adhere to refrigerators and other things. Other materials like iron and steel can be temporarily magnetized in the presence of

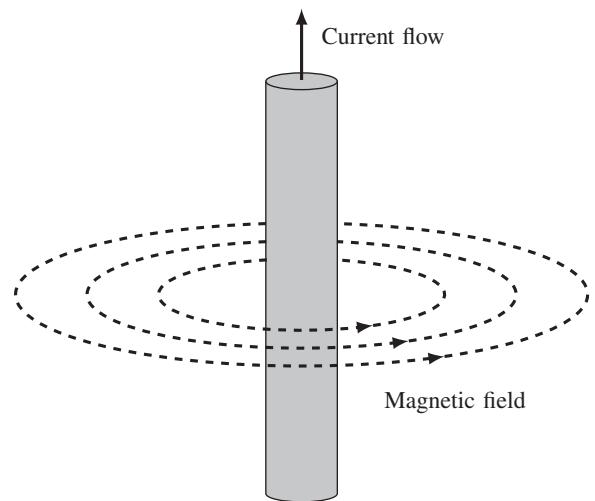


Figure 1.2 Magnetic field around a current-carrying wire.

a sufficiently strong external magnetic field (this is what happens to the refrigerator door underneath the magnet), with the particles returning to their disordered state after the external field is withdrawn.

The magnetic property of microscopic particles is due to their electric charge and their intrinsic motion, which brings us to the fundamental connection between electricity and magnetism. Indeed, we can think of magnetism as nothing but a manifestation of directionality associated with electric charge in motion, whereby moving charges always exert a specific directional force on other moving charges. At the level of individual electrons, their motion consists of both an orbital movement around the atom's nucleus and an intrinsic spin, which we can visualize as if the particle were spinning like a top.²⁶ Both of these rotational motions combine to form what is referred to as a magnetic moment. Similarly, the protons inside atomic nuclei possess a magnetic moment due to their intrinsic spin.²⁷

Knowing this, it would stand to reason that a large amount of moving charge such as a measurable electric current should produce a magnetic field as well. This phenomenon was in fact discovered in 1820, when Hans Christian Oersted observed that a compass needle was deflected by an electric current through a nearby wire. The magnetic field produced by an electric current points at a right angle to the flow of charge, in a direction specified by the “right-hand rule” illustrated in Figure 1.2. If the thumb of one’s right hand is pointing in the direction of the current, then the curled fingers of the same hand indicate the direction of the magnetic field. Thus, the magnetic field lines surround the wire in a circular manner.

In order to make practical use of this phenomenon, we can alter the shape of the current-carrying wire by winding it into a coil, which brings many turns of wire closely together so that their magnetic fields will add to form a “straight” field in the center of the coil that is comparable to that of a bar magnet (an illustration of such a coil and its magnetic field is shown in Figure 3.4). This arrangement can be thought of as “concentrating” the magnetic field in space.

²⁶ Such a mechanistic representation is not quite accurate in terms of quantum mechanics, but it is nonetheless useful for constructing some intuitive picture.

²⁷ This effect is quite subtle and not important in our context, but is exploited in such technologies as magnetic resonance imaging (MRI) for medical diagnostic purposes, which discriminates among tissues of different water content by way of the magnetic properties of the hydrogen nucleus.

Magnetic effects are essential for the generation and conversion of electric power. In order to successfully navigate the literature on these applications, it is important to be aware of a distinction between two types of quantities: one is called the *magnetic field* and the other *magnetic flux*. Despite the earlier caution, it is at times helpful (and indeed consistent with the Latin translation) to think of the flux as the directional “flow” of something, however immaterial, created in turn by the flow of electric current. Conceptually as well as mathematically, the flux is a very convenient quantity for analyzing electrical machines, while the magnetic field is particularly useful for describing the basic principles of electromagnetic induction in simplified settings.

Conventionally, the magnetic field is denoted by the symbol B and measured in units of *tesla* (T) or *gauss* (G). One tesla, which equals 10,000 or 10^4 gauss, corresponds to one newton (N, a measure of force) per ampere per meter: $1 \text{ T} = 1 \text{ N/A}\cdot\text{m}$. Magnetic flux is denoted by ϕ (the Greek phi) and is measured in units of *weber* (Wb). One tesla equals one weber per square meter.

From this relationship between the units of flux and field, we can see that the magnetic field corresponds to the density or concentration in space of the magnetic flux. The magnetic field represents magnetic flux *per unit area*. Stated in reverse, magnetic flux represents a measure of the magnetic field multiplied by the area that it intersects.

Unless “concentrated” by a coil, the magnetic field associated with typical currents is not very strong. For example, a current of 1 A produces a magnetic field of 2×10^{-7} T or 0.002 G (2 milligauss) at a distance of 1 m. By comparison, the strength of the Earth’s magnetic field is on the order of half a gauss.²⁸

1.5.4 Electromagnetic Induction

While electric current creates a magnetic field, the reverse effect also exists: magnetic fields, in turn, can influence electric charges and cause electric currents to flow. But there is an important twist: the magnetic field must be *changing* relative to the charge in order to have any effect. A static magnetic field, whether a bar magnet or the field surrounding a current-carrying wire, will not cause any motion of nearby charge—until there is relative motion between that charge and the magnetic field. Relative motion could mean that either the magnet or the wire is being moved through space; it could also mean that the strength of the magnetic field is changing. Whenever such movement or variation in the magnetic environment occurs, an electric charge will experience a force. The fundamental physical effect is called the *Lorentz force*. In the power systems context, where large numbers of electric charges reside inside conducting pieces of metal and are accelerated by a changing magnetic field, it is called the *electromotive force* or *emf*.

The simplest case of the Lorentz force involves a single charged particle traveling through a magnetic field in space, at a right angle to the magnetic field lines. The charge experiences a force again at right angles to both the field and its velocity, whose direction depends on the sign of the charge (positive or negative).

This effect can be expressed concisely in mathematical terms of a *cross product* of vector quantities (i.e., quantities with a directionality in space, represented in boldface), in what is known as the Lorentz equation,

$$\mathbf{F} = q \mathbf{v} \times \mathbf{B}$$

where \mathbf{F} denotes the force, q the particle’s charge, \mathbf{v} its velocity, and \mathbf{B} the magnetic field.²⁹ In the case where the angle between \mathbf{v} and \mathbf{B} is 90° (i.e., the charge travels at right angles to the direction

²⁸ The exact value of the Earth’s magnetic field depends on geographic location, and it is less if only the horizontal component (to which a compass needle responds) is measured.

²⁹ We are using boldface symbols for vectors; the charge q is just a number or *scalar* quantity.

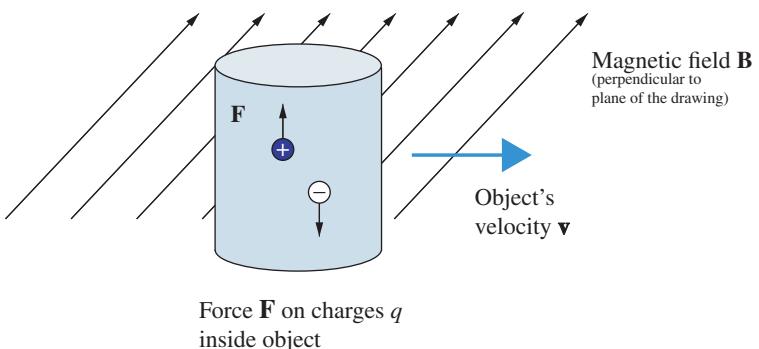


Figure 1.3 The Lorentz force acting on charges moving through a magnetic field.

of the field), the magnitude or numerical result for \mathbf{F} is simply the arithmetic product of the three quantities. This is the maximum force possible: as the term cross product suggests, the charge has to move across the field in order to experience the effect. The more \mathbf{v} and \mathbf{B} are at right angles to each other, the greater the force; the more closely aligned \mathbf{v} and \mathbf{B} are, the smaller the force. If \mathbf{v} and \mathbf{B} are parallel—that is, the charge is traveling along the magnetic field lines rather than across them—the force on the charge is zero.

Note that the vector cross product has a handedness, which can also be described as a “right-hand rule.” However, the physical process here is distinct from the one in Figure 1.2 where the magnetic field is being caused by moving charges. Instead, we are now describing the direction of force on a charge that is caused by its movement relative to a pre-existing magnetic field, just using the same hand.

Figure 1.3 illustrates a relevant application of this relationship for our context. We are looking at an object that contains electric charges. The entire object is moving from left to right, so that each of the microscopic charges inside has a velocity \mathbf{v} in the direction of motion. If we align our right hand with that direction \mathbf{v} and then curl our fingers in the direction of the magnetic field \mathbf{B} that points backward into the page, our thumb will point in the direction of the force \mathbf{F} on a positive charge (upward). A negative charge of the same magnitude experiences the opposite (downward) force.

Suppose the moving object in Figure 1.3 represents a piece of conducting metal wire. In a metal, only the negatively charged electrons are free to move, while the positive metal ions are stuck in place. The electromotive force would thus cause a downward movement of electrons. Realistically, though, this movement of electrons in an isolated piece of metal would stop very soon, since electrons would begin to accumulate near the bottom and therefore experience a mutual repulsion (which counteracts the *emf* from the motion in the magnetic field). What we would actually observe is a voltage or potential difference between the top and bottom of the object. However, if the wire were stretched out and connected into a complete circuit (as would be the case for a winding in an electric generator), then the electrons could continue moving and the *emf* would have caused a current to flow. Such a current is called an *induced current*, produced by the phenomenon of *electromagnetic induction*.

Because only the relative motion between the charge and the magnetic field matters, the same effect results if the charge or wire is stationary in space and the magnetic field is moved (e.g., by physically moving a bar magnet), or even if both the magnet and the wire are stationary but the magnetic field is somehow made to become stronger or weaker over time. As we will see in Chapter 10, a

combination of these effects—movement through space of wires and magnets, as well as changing magnetic field strength—is employed in the production of electric power with rotating generators.

1.5.5 Electromagnetic Fields and Health Effects

A current flowing through a wire, alternating at 60 cycles per second (60 Hz), produces around it a magnetic field that changes direction at the same frequency. Thus, whenever in the vicinity of electric equipment carrying any currents, our bodies are exposed to magnetic fields. Such fields are sometimes referred to as *EMF*, for *electromagnetic fields*, or more precisely as *ELF*, for *extremely low-frequency fields*, since 60 Hz is extremely low compared to other electromagnetic radiation such as radio waves (which is in the megahertz range).

There has been some concern in the scientific community that even fields produced by household appliances or electric transmission and distribution lines might present human health hazards. While such fields may be small in magnitude compared to the Earth's magnetic field, the fact that they are oscillating at a particular frequency could conceivably have biological implications that are as yet poorly understood.

Research on the health effects of EMFs or ELFs continues. Some results seemed to indicate a small but statistically significant correlation of exposure to ELFs from electric power with certain forms of cancer, particularly childhood leukemia, while other studies have found no effects.³⁰ In any case, the health effects of ELFs on adults appear to be either sufficiently mild or sufficiently rare that no obvious disease clusters have been noted among workers who are routinely exposed, and have been over decades, to vastly stronger fields (by orders of magnitude) than are commonly experienced by the general population.

From a purely physical standpoint, the following observations are relevant: First, the intensity of the magnetic field associated with a current in a wire is directly proportional to the current; second, the intensity of this field decreases at a rate proportional to the inverse square of the distance from the wire, so that doubling the distance reduces the field by a factor of about 4. The effect of distance thus tends to outweigh that of current magnitude, especially at close range where a doubling may equate to mere inches. Therefore, sleeping with an electric blanket or even an electric alarm clock on the bedside table would likely lead to much higher exposure than living near high-voltage transmission lines. Measured ELF data are published by many sources.

1.5.6 Electromagnetic Radiation

Although not vital in the context of electric power, another manifestation of electromagnetic interactions deserves at least brief discussion: namely, electromagnetic waves or *radiation*, including what we experience as light. Visible light in fact represents a small portion of an entire spectrum of electromagnetic radiation, which is differentiated by *frequency* or *wavelength*. The nonvisible (to us!) regions of this spectrum include infrared and ultraviolet radiation, microwaves, radio waves and others used in telecommunications (such as cellular phones), X-rays, and gamma rays from radioactive decay. Physically, all these types of radiation are of the same basic nature.

³⁰ Technical information and summaries of research have been published by the World Health Organization, <https://www.who.int/health-topics/electromagnetic-fields> and by the U.S. Environmental Protection Agency, <https://www.epa.gov/radtown/electric-and-magnetic-fields-power-lines> (accessed February 2024). See also C.J. Portier and M.S. Wolfe (eds.), *Assessment of Health Effects from Exposure to Power-Line Frequency Electric and Magnetic Fields*, NIEHS Working Group Report (Research Triangle Park, NC: National Institute of Environmental Health Sciences of the National Institutes of Health, 1998).

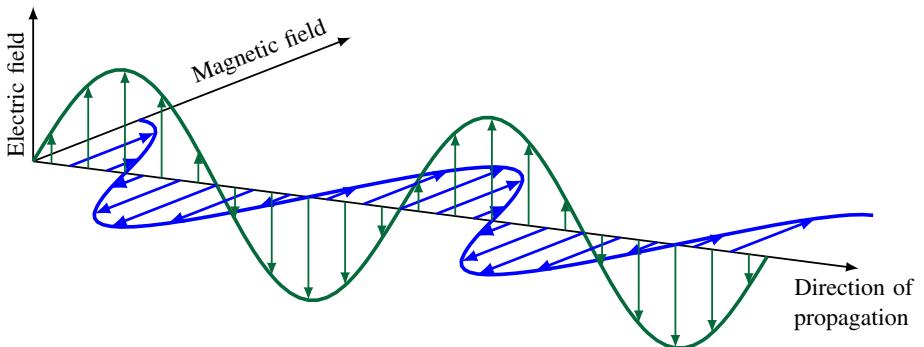


Figure 1.4 An electromagnetic wave.

The concept of a wave is familiar to most of us as the periodic movement of some material or medium: the surf on a Hawaiian beach, a vibrating guitar string, or the coordinated motion of sports fans in the bleachers. What is actually “waving” when electromagnetic radiation travels through space is much less tangible and challenging to the imagination; we can describe it only as a pulse of rapidly increasing and decreasing electric and magnetic fields, themselves completely insubstantial and yet measurably affecting their environment. These electric and magnetic fields are at right angles to each other, and at right angles again to the direction of propagation of the wave, as illustrated in Figure 1.4.

The *frequency* of the electromagnetic wave refers to the rate at which either the electric or magnetic field at any one point oscillates (i.e., changes direction back and forth). Frequency is measured in cycles per second, or *hertz* (Hz). The wavelength represents the distance in space from one wave crest to the next (analogously, the *period* of a wave represents the separation in time from one wave crest to the next). Depending on the range of the spectrum in question, wavelength may be measured in meters or any small fraction thereof. For example, the wavelength of a certain color of visible light might be quoted in microns ($1 \mu\text{m} = 10^{-6} \text{ m}$), nanometers ($1 \text{ nm} = 10^{-9} \text{ m}$), or ångströms ($1 \text{ \AA} = 10^{-10} \text{ m}$).

Wavelength is inversely proportional to the frequency; higher frequency implies shorter wavelength, and *vice versa*. This is because wavelength and frequency multiplied together yield the speed of propagation of the wave, which is fixed: the speed of light. Aside from the caveat that the speed of an electromagnetic wave actually varies slightly depending on the medium through which it is traveling, the constancy of the speed of light is famously important.³¹

This constancy can be understood as a manifestation of the principle of the *conservation of energy*: It is at this speed and only at this speed of propagation, or rate of change of electric and magnetic fields, that they keep inducing each other at the same magnitude. If the wave propagated more slowly, the fields would decay (implying energy that mysteriously vanishes); were it to propagate any faster, the fields would continually increase (implying a limitless creation of energy). As we know from the first law of thermodynamics, energy can be neither created nor destroyed in any physical process. From this basic principle, it is possible to derive the constant speed of light.

Electromagnetic radiation interacts with matter through charges—specifically, electrons—that are accelerated and moved by the field. Let us consider first the example of radio waves that are

³¹ In the fine print of the physics textbook we learn that $3 \times 10^8 \text{ m/s}$, or 186,000 mi/h, is the speed of light only in a vacuum. Light travels somewhat slower through various materials, and these small differences in the speed of light—both as a function of the medium and of wavelength—give rise to familiar optical phenomena like refraction in a lens, prism, or glass of water.

broadcast and received through conducting metal antennas, and then the more general case of photon absorption and emission in all types of materials.

As the music plays at the radio station, its specially encoded electronic signal travels in the form of a rapidly changing electric current into the station's large antenna, moving the electrons inside the metal up and down.³² These moving electrons produce a pulse of a changing electric field that is "felt" in the region of space surrounding the antenna. This oscillating electric field induces a magnetic field, which in turn induces an electric field, and so forth, with the fields propagating away in the form of a wave—an electromagnetic wave of a very specific time signature. The wave becomes weaker with increasing distance from the antenna in that it spreads out through space, though the "pulse" itself is preserved as long as the wave is detectable.

Another antenna at a distance can now "receive" the wave because the electrons inside it will be accelerated by the changing electric field, in exactly the same fashion as the electrons responsible for "sending" it. We can see that an antenna needs to be conducting so as to allow the electrons to move freely according to the changing field. When this induced motion of electrons is decoded by the radio receiver, the electric signal travels through a wire and finally moves the magnet of a loudspeaker back and forth, the specific signature of the electromagnetic wave is translated back into sound.

This large-scale motion of electrons, as in radio antennas, is a special case of their interaction with electromagnetic radiation. More generally, electrons stay within their atomic orbitals, but they nevertheless undergo certain transitions that allow them to "send out" or "receive" electromagnetic radiation. These transitions are not readily represented in terms of physical motion, but can only be described in the language of quantum mechanics. Physicists say that an electron changes its *energy level*, and that the difference between the energy level before and after the transition corresponds to the energy carried by a "packet" of electromagnetic radiation. Such a packet is called a *photon*.³³ As an electron moves to a state of lower energy, it emits a photon, and conversely, as it absorbs a photon, it rises to a state of higher energy.

Although the photon itself has no mass and can hardly be conceived of as an "object," it is nonetheless transporting energy through space. The amount of energy is directly proportional to the frequency of the radiation: the higher the frequency, the greater the energy. Thus, we can think of the "waving" electric and magnetic fields in space as a form of potential energy, whose presence does not become apparent until it interacts with matter.

The configuration of electrons within a given material, having a certain atomic and molecular structure, determines what energy transitions are available to electrons. They will interact with radiation only to the extent that the available transitions match precisely the energy of the photon, corresponding to its frequency (wavelength). This explains why materials interact differently with radiation of different frequencies, absorbing some and transmitting or reflecting others. A glass window, for example, transmits visible light but not ultraviolet. And we find ourselves—at this very moment!—in a space full of radio waves, oblivious to their presence because the waves pass right through our bodies, as the energy of their individual photons is insufficient to cause a transition

³² There are two standard types of encoding: amplitude modulation (AM) and frequency modulation (FM). In each case, the sound signal (which itself is an electrical pulse of changing voltage and current that mimics the corresponding sound wave) is superimposed on a carrier wave of a given frequency (the broadcast frequency of that particular station, which, at many kilohertz or megahertz, is several orders of magnitude higher than the frequency of the signal). In AM, the amplitude of the carrier wave is continually changed (modulated) according to the signal; in FM, the frequency is changed by a small percentage.

³³ It was one of the most stunning discoveries in early 20th-century physics that radiation occurs in such packets, or quanta, that only interact with a single electron at a time; the crucial experiment that demonstrated this (the photoelectric effect) is what actually won the 1921 Nobel Prize in Physics for Einstein.

of our electrons. To be sure, there are also photons of higher frequencies in our environment (ultraviolet from the sun, cosmic X-rays, and gamma rays) that occasionally *do* interact with our biochemistry, and chances are that any such interaction will have a deleterious effect on the cell.

In the context of electric power system operation, electromagnetic radiation does not play much of an explicit role. This is because the conventional frequency of alternating current at 50 or 60 Hz is so low that the corresponding radiation propagates with extremely little energy and is in practice unobservable. Stationary and alternating electric and magnetic fields, however, are central to the workings of all electric machinery.

Problems and Questions

- 1.1** An incandescent light bulb is rated 40 W at 120 V. What is its resistance in ohms, and what is the current? Repeat for a bulb rated 60 W.
- 1.2** Find the current and the power for a toaster oven supplied by 120 V, if the resistance is 12 or $24\ \Omega$, respectively. Explain in your own words why a load with lower resistance draws more power.
- 1.3** Suppose you accidentally touch a hot wire at 220 V. Assuming your shoes are the most significant resistance in the circuit, what must their resistance be so that the current through your body could not exceed 50 millamps (mA)?
- 1.4** Suppose a battery can deliver 100 amp-hours (Ah) at 12 V.
 - (a) How much stored energy does it contain, in kilowatt-hours? In joules?
 - (b) How much power in kW would be required to fully charge the battery in 20 minutes?
 - (c) If the battery were charged at 120 V at this rate, what is the current required? Ignore losses.
 - (d) Search online: what is the energy storage capacity in kWh of a regular Tesla Model 3 battery? How many AA batteries is that?
- 1.5** One Rosenfeld is defined as three billion kilowatt-hours per year, equivalent to the electricity production of a typical 500 megawatt (MW) coal-fired power plant operating at a capacity factor of roughly 68%.
 - (a) Knowing your scientific prefixes, and knowing how many hours there are in a year, infer the definition of the term “capacity factor” from the above sentence. (Note that capacity factor is distinct from efficiency, even though the numbers in this example might appear related.)
 - (b) Suppose this power plant converts thermal to electrical energy at an efficiency of one third (about 33%). What is the rate of waste heat production by the power plant in MW, and in kWh/year?
 - (c) Suppose the waste heat from one hour of operation of this power plant is absorbed by the water in an Olympic size swimming pool, $25 \times 50\text{ m}$, and 2 m deep, initially at a reasonable ambient temperature. Would that be enough heat to boil away the water in the pool?

- 1.6** A 100-W incandescent light bulb is replaced with a 13-W LED. If the light is on six hours every day, how many kilowatt-hours of energy are saved per month?
- 1.7** A laser printer is plugged into the wall outlet of a home office, 50 ft from the service panel where the voltage is 120.0 V. The wiring is 14-gauge copper (1.628 mm diameter, $\rho = 1.76 \times 10^{-8} \Omega\text{-m}$). When the printer heats up, it momentarily draws a starting current (called a “transient”) of 30 A.
- What is the voltage at the printer’s outlet during this transient?
 - By what fraction will this reduce the power to an incandescent lamp plugged in next to the printer? Would you expect to notice the light dimming momentarily?
- 1.8** Some energy efficiency experts have advocated using 12 AWG (American Wire Gauge) instead of 14 AWG copper as a standard in all home circuits. Estimate the reduction in power losses due to the increased wire size if there are 200 ft of wiring with a current of 10 A. Given this advantage, why do you think 12-gauge copper is not used everywhere?
- 1.9** Electrostatic discharge around the home and office—such as one experiences after scuffling across a carpet, and then touching a door knob or light switch—can be a nuisance and can damage sensitive electronic equipment.
- Given that dry air will support an electric field of $3 \times 10^6 \text{ V/m}$, what must be your body’s voltage relative to ground if you draw a visible arc when your finger is a millimeter away from the screw on a light switch plate?
 - Explain in your own words why this experience, however unpleasant, does not pose a danger to your health.
 - Suppose the spark delivers 1 mJ of energy. What other information would you need to calculate the current in amps? Make an assumption and estimate.
- 1.10** The material used to insulate electrical wires is characterized in volts per mil (1 mil = 0.001 in.) that it can sustain. Numbers on the order of 100 V/mil are typical. Suppose a certain 14-gauge wire rated for 240 V has an outer diameter (including insulation) of 0.11 in., while the uninsulated copper wire has a diameter of 0.07 in.
- What is the thickness of the insulation layer in mils and how many volts per mil must it sustain under normal conditions?
 - Explain why this number is, and should be, significantly smaller than the typical material rating.
- 1.11** A *muon* is an elementary particle with the same charge as an electron, but a much greater mass. Muons come from cosmic rays interacting with the earth’s upper atmosphere. Suppose a muon travels at $0.9 c$ (where the speed of light $c = 3 \times 10^8 \text{ m/s}$) directly downward from space. It happens to traverse a horizontal magnetic field of 3 T inside an MRI machine.
- Draw a sketch of the muon’s velocity, the field, and the resulting force on the muon.
 - Calculate the force in units of newtons. (This is a basically a unit conversion exercise.)
 - How do you visualize the path taken by the muon? How would the muon’s path differ from that of an electron traveling at the same velocity?
- 1.12** A radio station broadcasts a signal with a carrier wave of frequency 88.5 MHz. What is the wavelength?

1.13 With a bit of online research, identify a typical range or order of magnitude of electric power use in watts (kW, MW, GW) for the following. You may indicate either peak or average quantities, but state clearly which they are.

- (a) Laptop
- (b) Refrigerator
- (c) Electric vehicle
- (d) Single-family home
- (e) A building on your university campus
- (f) Your university campus
- (g) Your city
- (h) Your state or country
- (i) The world

2

DC Circuit Analysis

This chapter introduces foundational concepts of electric circuits, involving only *direct current* (DC or d.c.). All concepts in this chapter—series and parallel connections, Kirchhoff's laws, superposition, and equivalent circuits—apply equally to circuits with alternating current, but are illustrated here with d.c. examples in the interest of learning one new thing at a time.

2.1 Modeling Circuits

As a general definition, a circuit is an interconnection of electric devices, or physical objects that interact with electric voltages and currents in a particular manner. Typically, we would imagine the devices in a circuit to include a power source (such as a battery, a wall outlet, or a generator), conductors or wires through which the electric current can flow, and a *load*. The term “load” can refer to any circuit element, device, or combination of devices where electric power is being extracted from the circuit, for example, converted to mechanical or thermal energy. To analyze a circuit means to account for the properties of all the individual devices so as to predict the circuit’s electrical behavior. By “behavior” we mean specifically what voltages and currents will occur at particular places in the circuit given some set of conditions, such as a voltage supplied by a power source. This behavior will depend on the nature of the devices in the circuit and on how they are connected.

For the purpose of circuit analysis, individual devices are represented as ideal objects or circuit elements that behave according to well-understood rules.¹ From the circuit perspective, the events inside these elements are irrelevant; rather, we focus on measurements at the elements’ terminals, or points where the elements connect to others.

The scale of analysis can shift depending on the information that is of interest. For example, suppose we are analyzing a circuit in our house to see whether it might be overloaded. We find that several appliances are plugged into this circuit, including a radio. We would consider this radio as one of the loads and conceptualize it simply as a box with two terminals (the two prongs of its plug) that draws a particular amount of current when presented with 120 volts by the outlet. On the other hand, suppose we wish to understand how the radio can be tuned to different frequencies. Now we would draw a diagram that includes many of the radio’s interior components. Still, we might not include every single little resistor or capacitor; rather, we would group some of the many electronic

¹ Most real devices match their simple idealized versions very closely in behavior (e.g., a resistor that obeys Ohm’s law). If not, there is usually some way of combining a set of abstract elements so as to represent the behavior of the physical gadget to the desired accuracy.

parts together and represent them as a single element so as to eliminate unnecessary detail. Our choice of scale for this grouping would be precisely such that what goes on inside the elements we have defined is not relevant to the question at hand.

Grouping components into functional elements on whatever the appropriate scale is a powerful technique in circuit analysis, and electrical engineers use it constantly without even thinking about it. As we will see, there are precise rules for “scaling up” or simplifying the representation of a circuit without altering the relevant properties.

The most basic circuit elements, and those mainly of interest in power systems, include resistors, capacitors, and inductors. These are also called linear circuit elements because they exhibit linear relationships between voltage and current or their rates of change. In electronics, nonlinear circuit elements such as transistors and diodes are also extremely important, but they are beyond the scope of this chapter.² Finally, common circuit elements include d.c. and a.c. power sources, some of which may be broken down further for analysis in certain contexts (e.g., the internal modeling of electric generators).

Circuit models are generally drawn at a level of abstraction such that the conductors connecting various circuit elements can be assumed to have a negligible resistance. Thus, they are simply drawn as solid lines, of arbitrary length and shape, with no particular significance other than the endpoints they connect. For purposes of circuit analysis, two points connected by such a wire might as well be immediately adjacent; they are, in electrical terms, the same point.³ Formally, we would say that the voltage difference between any two points connected by a zero-resistance conductor is nil; the points are at the same electric potential.

A notable distinction in the power systems context as compared to smaller-scale circuits is that when we are dealing with transmission and distribution lines that extend over longer distances, the assumption of negligible resistance (or, more accurately, impedance) for a conductor no longer holds. Thus, depending again on the scale of analysis, conductors in power systems often need to be represented as explicit circuit elements in their own right.

2.2 Series and Parallel Circuits

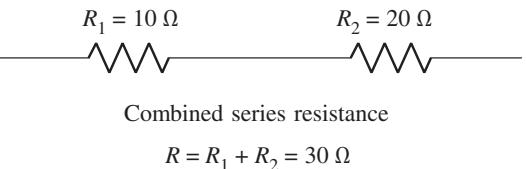
When considering multiple devices in a circuit and their joint behavior, it is obviously important how they are connected. There are two basic ways in which circuit devices can be connected together, referred to as *series* and *parallel*.

A series connection is one in which the electric current flows first through one element, then through the next. By necessity, all the current that goes through the first also goes through the second (and third, etc.); in other words, the current through elements in series is equal. This requirement follows intuitively from the nature of current as a flow of charge. This charge is neither created nor destroyed at any point along the connection of interest; we say it is *conserved*. Therefore, what goes in one end must come out of the other.

A parallel connection is one in which the devices present the current with two or more alternate paths: there is a branch point. Any individual charge will only go through one device,

² As discussed further in Chapter 14, nonlinear circuit elements have conductive properties that are not fixed (as we assume for all linear circuit elements), but that can change depending on ambient conditions (such as the voltage that is being applied, or the presence of a current). Usually, these circuit elements are made from combinations of semiconductor materials. Simpler resistors, capacitors, and inductors can also be nonlinear, but we will ignore those cases here.

³ Mathematically speaking, we are only modeling the topology of the connections.

Figure 2.1 Resistors in series.

and thus the current flow divides. Conservation of charge in this case requires that the sum of currents through all the alternate paths remain constant; that is, they add together to equal the initial current that was divided up. This notion is formalized in Kirchhoff's current law (KCL) (see Section 2.3.2).

Any network of circuit elements, no matter how intricate, can be decomposed into series and parallel combinations.

2.2.1 Resistance in Series

The simplest kind of combination of multiple circuit elements has resistors connected in series (Figure 2.1). The rule is easy: to find the resistance of a series combination of resistors, add their individual resistances. For example, if a 10-V resistor is connected in series with a 20-V resistor, their combined resistance is 30 V. This means that we could replace the two resistors with a single resistor of 30 V and make no difference whatsoever to the rest of the circuit. In fact, if the series resistors were enclosed in a box with only the terminal ends sticking out, there would be no way for us to tell by electrical testing on the terminals whether the box contained a single 30-V resistor or any series combination of two or more resistors whose resistances added up to 30 V. Thus, an arbitrary number of resistances can be added in series, and their order does not matter.

Intuitively, the addition rule makes sense because if we think of a resistor as posing an “obstacle” to the current, and note that the same current must travel through each element in the series, each obstacle adds to the previous ones. This notion can be formalized in terms of *voltage drop* (defined in Section 1.3.3). Across each resistor in a series combination, there will be a voltage drop proportional to its resistance. It is always true that, regardless of the nature of the elements (whether they are resistors or something else), the voltage drop across a set of elements connected in series equals the sum of voltage drops across the individual elements. This notion reappears in the context of Kirchhoff's voltage law (KVL) (see Section 2.3.1).

2.2.2 Resistance in Parallel

When resistors are combined in parallel, the effect is perhaps less obvious than for the series case: rather than adding resistance, we are in fact *decreasing* the overall resistance of the combination by providing alternative paths for the current. This is so because in the parallel case, the individual charge is not required to travel through every element, only one branch, so that the presence of the parallel elements “alleviates” the current flow through each branch, and thereby makes it easier for the charge to traverse. It is convenient here to consider resistors in terms of the inverse property, *conductance* (Section 1.2.2). Thus, we think of the resistor added in parallel not as posing a further obstacle, but rather as providing an additional conducting option: after all, as far as the current is concerned, any resistor is still better than no path at all. Accordingly, the total resistance of a parallel combination will always be *less* than any of the individual resistances.

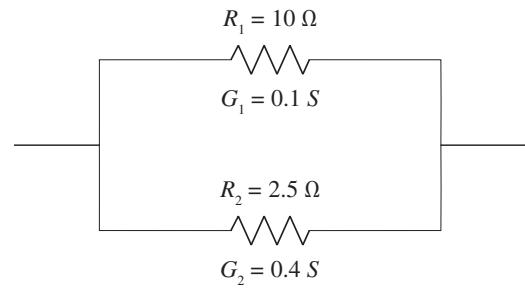


Figure 2.2 Resistors in parallel.

Combined parallel resistance

$$R = R_1 R_2 / (R_1 + R_2) = 2\Omega$$

$$G = G_1 + G_2 = 0.5 S$$

Using conductance ($G = 1/R$), the algebraic rule for combining any number of resistive elements in parallel is simply that the conductance of the parallel combination equals the sum of the individual conductances.⁴

For example, suppose a 10Ω and a 2.5Ω resistor are connected in parallel, as in Figure 2.2. We know already that their combined (parallel) resistance must be less than 2.5Ω . To do the math, it is convenient to first write each in terms of conductance: $0.1 S$ and $0.4 S$. The combined conductance is then simply the sum of the two, $0.5 S$. Expressed in terms of resistance, this result equals 2Ω . In equation form, we would write for resistors in parallel:

$$\frac{1}{R} = \frac{1}{R_1} + \frac{1}{R_2} \dots$$

where R is the combined resistance, and R_1 , R_2 , and so forth are the individual resistances.

In the case of only two resistances in parallel, it is often preferred to rearrange this equation to read⁵

$$R = \frac{R_1 R_2}{R_1 + R_2}$$

With more than two resistances, however, this notation becomes increasingly awkward. For example, in the case of three resistances in parallel, it reads

$$R = \frac{R_1 R_2 R_3}{R_1 R_2 + R_1 R_3 + R_2 R_3}$$

Thus, when many resistances are combined in parallel, it is generally more convenient to express them in terms of conductance.

Note that the voltage drop across any number of elements in parallel is the same. This can easily be seen because all the elements share the same terminals: the points where they connect to the rest of the circuit are, in electrical terms, the same.

While elements connected in parallel thus have a common voltage drop across them, the current flowing through the various elements or branches⁶ will typically differ. Intuitively, we might guess

⁴ Common circuits do not include superconductors, so mathematicians are asked not to worry about the case where $R = 0$. The superconducting situation is briefly addressed in Section 1.2.2, but analyzing it quantitatively requires techniques beyond the scope of this book.

⁵ When memorizing this formula, it is helpful to keep in mind that the units of both sides of the equation are resistance. Thus, the product term (units of resistance squared) must be in the numerator, and the sum term (units of resistance) in the denominator.

⁶ The term “branch” for one path in a parallel connection is preferable because, in general, there could be more than one element (in series) along each parallel path; see Section 2.2.5.

that more current will flow through a branch with a lower resistance, and less current through one with a higher resistance. This can be shown rigorously by applying Ohm's law for each of the parallel resistances: If V is the voltage drop common to all the parallel resistances, and R_1 is the individual resistance of one branch, then the current I_1 through this branch is given by V/R_1 . Thus, the amount of current through each branch is inversely proportional to its resistance.

For example, in Figure 2.2, the current through the $2.5\text{-}\Omega$ resistor will be four times greater than that through the $10\text{-}\Omega$ resistor, whatever the applied voltage. If the voltage is, say, 10 V, then the currents will be 1 and 4 A, respectively. Note that the sum of these currents is consistent with applying Ohm's law to the combined resistance: $10\text{ V}/2\ \Omega = 5\text{ A}$.

To summarize, there is a tidy correspondence between the series and parallel cases: In a series connection, the current through the various elements is the same, but the voltage drops across them vary (proportional to their resistance); in a parallel connection, the voltage drop across the various elements is the same, but the currents through them vary (inversely proportional to their resistance). These important observations remain true when we introduce alternating-current behaviors in Chapter 3, substituting the more general complex *impedance* for the resistance.

2.2.3 Network Reduction

As stated earlier, any network of circuit elements is composed of some mixture of series and parallel combinations. To analyze the network, circuit branches are sequentially aggregated from the bottom up as series or parallel combinations, up to the desired scale. The point is best made through an illustrative example.

Suppose we wish to determine the equivalent resistance of the network of five resistors shown in Figure 2.3, relying on the fact that any parallel and series combination of resistors can ultimately be reduced to a single resistance. Starting from the largest scale, we note that the total resistance will be the sum of R_5 and the combination of four resistors on the left-hand side. This combination, in turn, has a resistance corresponding to the parallel combination of R_4 and the branch on top with three resistances, and so on. For computation, we begin at the smallest scale, evaluating first the combination of R_1 and R_2 and working our way up from there. As demonstrated in the numerical example, it is convenient to switch back and forth between units of resistance and conductance so as to facilitate the arithmetic for evaluating the parallel combinations. Finally, by expressing the five individual resistances as one combined resistance, we have converted our initial model into a much simpler one with only a single circuit element, while its behavior in relation to anything outside it remains unchanged. Thus, we have effectively scaled up our

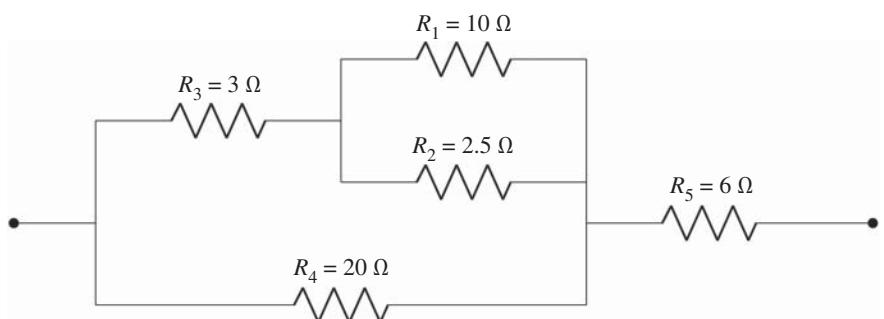


Figure 2.3 Network reduction.

representation of the circuit. This kind of simplification process can be carried out repeatedly at various levels.

Example

What is the combined resistance of the five resistors in the network in Figure 2.3?

We proceed by adding together conductances for parallel elements, and resistances for series elements, starting with the combined conductance of R_1 and R_2 , which is $G_{12} = G_1 + G_2 = 0.5 \text{ S}$. We then invert G_{12} to express it as a resistance that can be added to R_3 . Thus, $R_{12} = 1/G_{12} = 2\Omega$, and $R_{123} = R_{12} + R_3 = 5\Omega$. We then invert to conductance, $G_{123} = 1/R_{123} = 0.2 \text{ S}$, and add it to G_4 to obtain $G_{1234} = G_{123} + G_4 = 0.25 \text{ S}$. Finally, after inverting again to resistance, $R_{12345} = R_{1234} + R_5 = 10\Omega$.

In analyzing power systems, it is often necessary to model the system at different scales, depending, for example, on whether the focus is on long-distance transmission or local power distribution. The scaling process primarily involves aggregating individual loads within an area and representing them as a single block of load. This could be appliances in a house, the load on a distribution transformer, a distribution feeder, or an entire substation. In this case, however, the simplification is based on empirical measurement (e.g., of loads at the substation), as opposed to an algebraic procedure like the one just discussed, say, by combining the individual resistances or impedances (see Section 3.3).

In general, circuit analysis, and especially the technique of network reduction, gets more complicated when there is a large number of branches and circuit elements, and also when there are circuit elements of different types whose characteristics are not readily summarized in terms of resistance or impedance. For these situations, electrical engineers carry an arsenal of more sophisticated reduction techniques. The most important of these involve so-called Thévenin and Norton equivalent circuits, discussed in Section 2.5 below. As we will see in Chapter 12, the complexity of power systems is such that even small networks become quickly intractable by pencil and calculator. In that case, the analyst's task is to assemble all the information about the network neatly in tabular or *matrix* form, and then pass the arithmetic to a computer. We postpone this topic until Chapter 12.

2.2.4 Dual Concepts

It may seem intuitive to the reader that series and parallel combinations are somehow complementary opposites of each other, in a similar sense as conductance is the inverse or opposite of resistance. In electrical circuit analysis, these are called *dual* concepts. The beauty of dual concepts is that a statement about a circuit implies a statement about a dual situation, wherein each individual component or aspect is converted to its dual, and the statement holds true collectively. For example, we can say that [resistances/conductances] in [series/parallel] are additive. Each of the new concepts introduced in the subsequent Section 2.3 will have a dual.

2.2.5 Practical Aspects

In real circuits for power delivery, we mostly think of circuit elements in terms of power sources or loads, such as appliances in the house or several homes on the distribution circuit that runs down the block. These loads are always connected in parallel, not in series. The reason for this is that a

parallel connection essentially allows each load to be operated independently of the others, since each is supplied with the same standard voltage but can draw a current depending on its particular function (which determines the amount of power consumed).

Supplied by a constant voltage source, which we like to assume in the context of power systems (although it is only an approximation), resistive loads in parallel are essentially unaffected by each other. Interactive effects only occur as departures from the idealized situation; for example, when a particularly heavy load affects⁷ the local voltage, and thus indirectly the other loads. By contrast, independent operation of loads in series would be impossible, since elements in series share the same current. Thus, turning any one of them off will interrupt the current flow to all the others.⁸ Even if all elements are operational, the amount of power consumed in each one cannot easily be adjusted, and the voltage across each represents only some fraction of the voltage across the combination.

The series connection is mostly relevant in power systems when elements are considered that represent successive steps between power generation and consumption. Thus, the loads are in series with a distribution line, a transmission line, and a generator. Actually, these elements may also be in parallel with many others, since the entire system forms a network with many links. Still, there is some minimum number of elements in series that constitute a path from power source to load. For example, if we are interested in the resistance of the conductors between a distribution substation and a customer (where there is usually only a single path due to the radial layout of the distribution system; see Section 7.1.5), we would have to add all the contributions to resistance along the way.

The important conceptual point here is that because of this series connection, there is no escaping the interdependence among the elements: there is literally no way around the other elements on a series path. This becomes important in the context of transmission constraints (see Section 7.3) and excessive voltage drops due to high loads (Sections 1.3.3, 7.4, and 8.6). While it may be perfectly obvious to an engineer that any devices through which the same current must travel are necessarily dependent on each other, this presents a very fundamental problem when legal and institutional arrangements concerning power systems have these devices under the auspices of different parties.

2.3 Kirchhoff's Laws

Anything we learn about the behavior of a circuit from the connections among its elements can be understood in terms of two constraints known as Kirchhoff's laws.⁹ Specifically, they are KVL and KCL. Their application in circuit analysis is ubiquitous, sometimes so obvious as to be done unconsciously, and sometimes surprisingly powerful. While Kirchhoff's laws are ultimately just concise statements about the basic physical properties of electricity discussed in Chapter 1, when applied to intricate circuits with many connections, they turn into sets of equations that organize our knowledge about the circuit in an extremely elegant and convenient fashion.¹⁰ KVL and KCL are dual concepts.

⁷ Dimming or flickering lights when a big motor switches on are a familiar example.

⁸ Holiday lights of older vintages are a classic example of this phenomenon, and of the painstaking process to identify the culprit element. Newer models have a contact that allows the current to bypass the individual light bulb if it is broken, and the other lights on the string will only go out if a bulb is removed altogether.

⁹ After the 19th century German physicist Gustav Robert Kirchhoff, whose series combination of consonants is essentially unpronounceable for native English speakers. Luckily, the approximations KIRK-off or KERCH-off are equally recognizable.

¹⁰ In mathematical terms, Kirchhoff's laws yield a number of linearly independent equations, often arranged in matrix form, which are just sufficient to determine the voltages and currents in every circuit branch, given information about all the circuit elements present.

2.3.1 Kirchhoff's Voltage Law

Kirchhoff's voltage law (often abbreviated KVL) states that the sum of voltages around any closed loop in a circuit must be zero. In essence, this law expresses the basic properties that are inherent in the definition of the term “voltage” or “electric potential.” Specifically, it means that we can definitively associate a potential with a particular point that does not depend on the path by which a charge might get there. This also implies that if there are three points (A , B , and C) and we know the potential differences between two pairings (between A and B and between B and C), this determines the third relationship (between A and C). Without thinking in such abstract and general terms, we apply this principle when we move from one point to another along a circuit by *adding* the potential differences or voltages along the way, so as to express the *cumulative* voltage between the initial and final point. Finally, when we go all the way around a closed loop, the initial and final point are the same, and therefore must be at the same potential: a zero difference in all.

The analogy of flowing water comes in handy. Here, the voltage at any given point corresponds to the elevation. A closed loop of an electric circuit corresponds to a closed system like a water fountain. The voltage “rise” is a power source—say, a battery—that corresponds to the pump. From the top of the fountain, the water then flows down, maybe from one ledge to another, losing elevation along the way and ending up again at the bottom. Analogously, the electric current flows “down” in voltage, maybe across several distinct steps or resistors, to finish at the “bottom” end of the battery. This notion is illustrated by in the simple circuit in Figure 2.4 that includes one battery and two resistors. Note that it is irrelevant which point we choose to label as the “zero” potential: no matter what the starting point, adding all the potential gains and drops encountered throughout the complete loop will give a zero net gain.

2.3.2 Kirchhoff's Current Law

KCL states that the currents entering and leaving any branch point or node in the circuit must add up to zero. This follows directly from the conservation property: electric charge is neither created nor destroyed, nor is it “stored” (in appreciable quantity) within our wires, so that all the charge that flows into any junction must also flow out. Thus, if three wires connect at one point, and we know the current in two of them, they determine the current in the third.

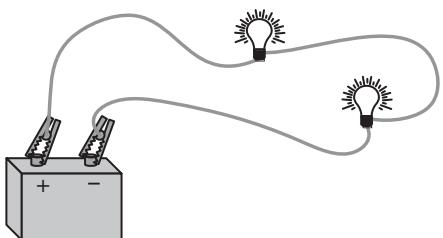


Figure 2.4 Kirchhoff's voltage law.

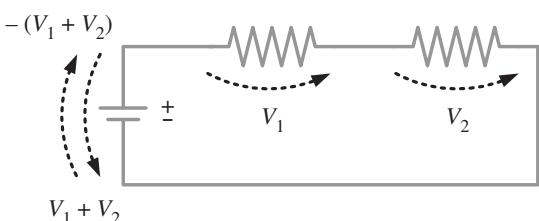
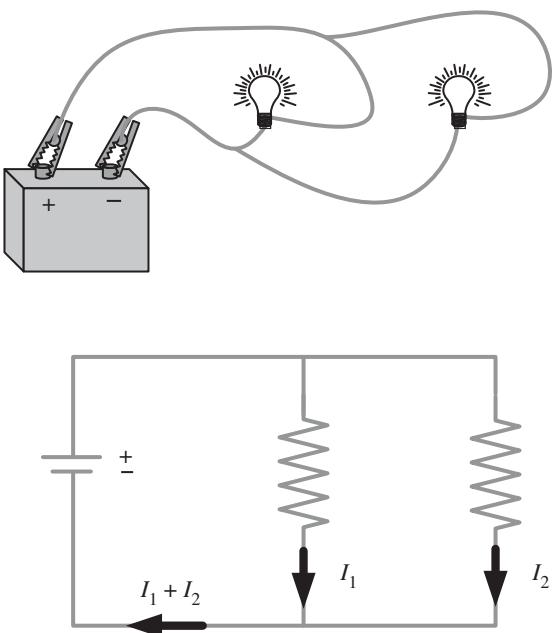


Figure 2.5 Kirchhoff's current law.



Again, the analogy of flowing water helps make this more obvious. At a point where three pipes are connected, the amount of water flowing in must equal the amount flowing out (unless there is a leak). For the purpose of computation, we assign positive or negative signs to currents flowing in and out of the node, respectively. It does not matter which way we call positive, as long as we remain consistent in our definition. Then, the sum of currents into (or out of) the node is zero. This is illustrated with the simple example in Figure 2.5, where KCL applied to the branch point proves that the current through the battery equals the sum of currents through the individual resistors.

Despite their simple and intuitive nature, the fundamental importance of Kirchhoff's laws cannot be overemphasized. They lie at the heart of the interdependence of the different parts and branches of power systems: whenever two points are electrically connected, their voltages and the currents through them must obey KVL and KCL, whether this is operationally and economically desirable or not. For example, managing transmission constraints in power markets is complicated by the fact that the flow on any one line cannot be changed independently of others. Thus, the engineer's response to the economist's lamentation of how hard it is to manage power transmission: "Blame Kirchhoff."

2.3.3 Application to Simple Circuits

When Kirchhoff's laws are combined with information about the characteristics of all the elements within a circuit (which specify the relationship between the voltage across and the current through each element), the voltages and currents at every location in a circuit can be specified regardless of the number of branch points, as long as all the circuit elements behave according to certain rules and the topology (connectedness) of the circuit meets certain criteria. Actually carrying out such a calculation, however, can be quite tedious, and we will only give a qualitative description of the process here.

The technique of choice is to write the relationships implied by Kirchhoff's laws as a list or table of equations. Such a table, when properly organized, is known as an *admittance matrix* (referring

to admittance as the generalized version of conductance in a.c. circuits). The admittance matrix shows which branch point or node of the circuit is connected to which other node, and also includes information about each branch between pairs of nodes. For large circuits, one writes this matrix in a systematic fashion by first labeling nodes and assigning reference directions for voltages between nodes and for the current through each branch (since it is important whether the voltage and current in any one direction is to be considered positive or negative). When combined according to the rules of linear algebra with current and voltage *vectors*—which are just properly ordered listings of all the branch currents and voltages—this formulation spells out Kirchhoff's laws. It is then a matter of standard linear algebra procedure to solve for any unknown current and voltage variables, given a sufficient number of “knowns.” This type of procedure is applied in Chapter 12 on power flow analysis. For now, let us illustrate some simple, familiar applications of Kirchhoff's laws through two examples.

Example

A string of holiday lights that plugs into a 120-V outlet has 50 identical bulbs connected in series. What is the voltage across each bulb?

The voltage across each bulb is $120\text{ V}/50 = 2.4\text{ V}$. This may seem obvious, but can be understood as a consequence of KVL, which requires that the voltage drops along the string add up to the same amount as the voltage drop from one to the other terminal of the wall outlet.

Example

A college student sets up a kitchen in his studio and decides to run several appliances from the same extension cord, with a multiplug at the end: a hot plate with an internal resistance of 20Ω , a light with 60Ω , and a toaster oven with 10Ω . The extension cord has a resistance of 0.1Ω on each of its two wires. What is the current through each device, when all are in use?

To a first approximation, we would say that the appliances all “see” a voltage of 120 V . The current through each is then given by Ohm's law: $120\text{ V}/20\Omega = 6\text{ A}$, $120\text{ V}/60\Omega = 2\text{ A}$, and $120\text{ V}/10\Omega = 12\text{ A}$.

If we care about greater accuracy, however, we should take into account the fact that at the end of the extension cord, the voltage is not exactly 120 V . According to KCL, the current through the cord must equal the sum of all three, or $6\text{ A} + 2\text{ A} + 12\text{ A} = 20\text{ A}$. To get a sense of how significant a difference this makes, we can quickly check that at 20 A , there would be a voltage drop of $20\text{ A} \cdot 0.1\Omega = 2\text{ V}$ on each “leg” of the cord. From KVL, it follows that the actual voltage at the end of the cord would then be only $120\text{ V} - (2 \cdot 2\text{ V}) = 116\text{ V}$. (This voltage is seen equally by all three appliances, which are connected in parallel through the multiplug.) The roughly 3% discrepancy may motivate us to calculate an exact answer.

For this, we write an expression for the total (series and parallel) resistance in the circuit:

$$R = 0.1 + \frac{1}{\frac{1}{20} + \frac{1}{60} + \frac{1}{10}} + 0.1 = 6.2\Omega$$

The current in the cord is thus $120\text{ V}/6.2\Omega = 19.35\text{ A}$. This current results in a voltage drop of twice $19.35\text{ A} \cdot 0.1\Omega$, leaving $120 - 3.87\text{ V} = 116.13\text{ V}$ for the appliances. The current in each appliance then comes out as 5.81, 1.94, and 11.6 A, respectively, to three significant figures. (Note that the three currents add up to the total current through the cord, as required by KCL.) Under most practical circumstances, neither the resistances nor the outlet voltage would really be known accurately enough to justify this calculation to three significant figures.

The moral of this story is that one ought to be careful about using extension cords, especially when they are long and of narrow gauge (high resistance) and when using powerful appliances. Some appliances—not the resistive kind in this example, but those involving motors—will run less efficiently and may eventually even be damaged when supplied with a voltage much less than the nominal 120 V. Worse yet, the extension cord may become very hot and pose a fire hazard.

2.4 The Superposition Principle

In addition to applying Kirchhoff's laws and scaling circuits up or down, a third analysis tool is based on the *superposition principle*. This principle applies to circuits with more than one voltage or current source. It states that the combined effect—that is, the voltages and currents at various locations in the circuits—from the several sources is the same as the sum of individual effects. Knowing this allows one to consider complicated circuits in terms of simpler components and then combining the results.

In power systems, the superposition principle is used to conceptualize the interactions among various generators and loads. For example, we could think of the current or power flow along a transmission link due to a “shipment” from one generator to one consumer, and we add to that the current resulting from separate transactions in order to obtain the total flow on that link.¹¹ With voltages held fixed, the currents become synonymous with power flows, and we can add and subtract megawatt flows superimposed along various transmission links. This procedure is illustrated in Section 7.1.7 on loop flow.

For a simple example where we can deal with currents and voltages explicitly, consider the circuit in Figure 2.6. This circuit has two power sources. The first source, labeled S_1 , is a battery that functions as a *voltage source*, delivering 12 V. The second source, S_2 , is a current source that delivers 1.5 A. This type of source may seem less familiar; it has the property of always delivering a specific current, regardless of the resistance in the circuit connected to it, while allowing the voltage across its terminals to vary.¹²

Suppose we wish to predict the voltage level v and current i at the locations identified in the diagram. It is not immediately obvious what the voltages and currents at various points in this circuit should be as a result of the combination of the two sources. The superposition principle is an indispensable analytic tool here: it states that we can consider separately the voltage and current that would result from each individual source, and then simply add them together. This principle applies regardless of the circuit's complexity or the number of power sources; it also holds true at any given instant in a circuit with time-varying sources.

In our example, the voltage v across R_2 that would result from only S_1 —written as $v(S_1)$, indicating that v for now is only a function of S_1 —is determined from the relative magnitudes of the two resistances in the circuit, R_1 and R_2 , while ignoring the presence of the current source. Since R_2 at 2Ω represents one-third of the total resistance in this simple series circuit, $2\Omega + 4\Omega$, the voltage across R_2 that we want to find is simply one-third of the total:

$$v(S_1) = 12 \text{ V} \cdot 2\Omega / (4\Omega + 2\Omega) = 12 \text{ V} \cdot 1/3 = 4 \text{ V}$$

¹¹ Power or energy are not, in fact, physically transported across the electric grid as distinct “shipments” from one place to another. This conceptualization is a crutch to convey the idea that the effects of multiple actors are overlaid and additive.

¹² Although voltage sources are more common in power systems, we introduce the current source here because it makes for a good illustration of the superposition principle; see more about sources in Section 2.5.

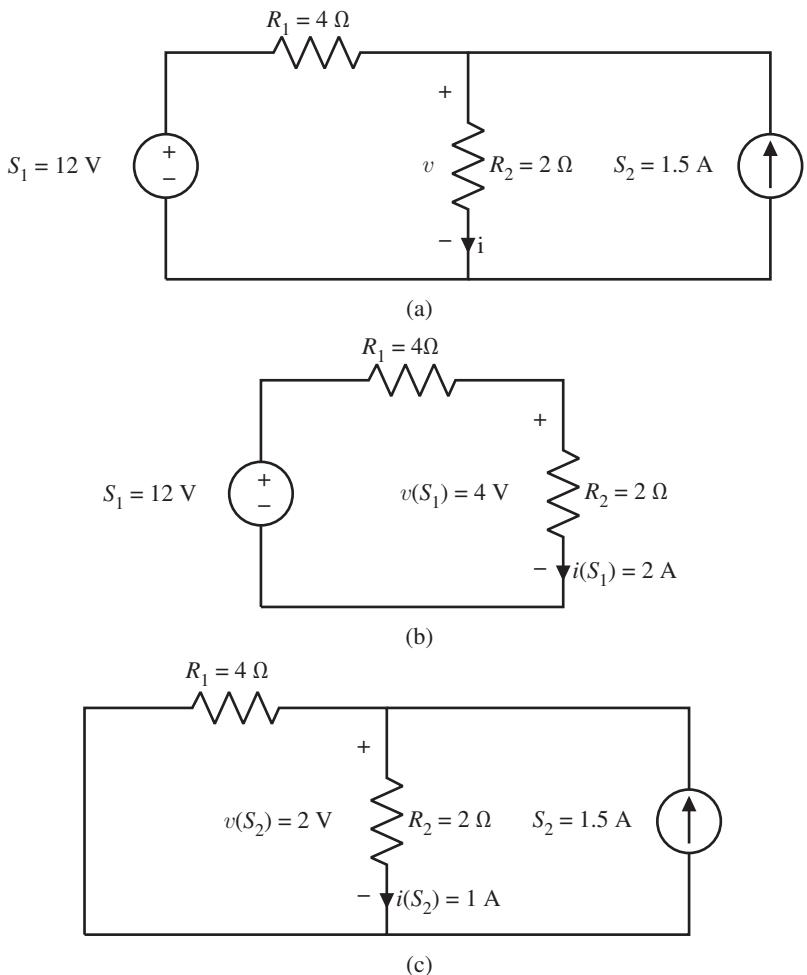


Figure 2.6 Superposition. (a) Full circuit. (b) Circuit with only the voltage source. (c) Circuit with only the current source.

Ignoring the current source technically means setting the current through it to zero, or replacing it with an open circuit as in Figure 2.6b. Having gotten rid of the extra circuit branch, the current $i(S_1)$ through the resistor R_2 that would result from S_1 alone is easy to find with Ohm's law:

$$i(S_1) = 12 \text{ V} / (4 \Omega + 2 \Omega) = 2 \text{ A}$$

Next, we ignore the voltage source, meaning that we set the voltage difference across it to zero, or replace it with a short circuit as in Figure 2.6c. The current $i(S_2)$ through R_2 based on the current source alone is found from the relative magnitudes of the resistances in each branch: since R_2 has half the resistance of R_1 , twice the current will flow through it, or two thirds of the total:

$$i(S_2) = 1.5 \text{ A} \cdot 4 \Omega / (4 \Omega + 2 \Omega) = 1.5 \text{ A} \cdot 2/3 = 1 \text{ A}$$

The voltage $v(S_2)$ due to the current source alone is again found via Ohm's law:

$$v(S_2) = 1 \text{ A} \cdot 2 \Omega = 2 \text{ V}$$

We can now superimpose the contributions from the two sources and find

$$v(S_1) = v(S_1) + v(S_2) = 4\text{ V} + 2\text{ V} = 6\text{ V}$$

and

$$i(S_1) = i(S_1) + i(S_2) = 2\text{ A} + 1\text{ A} = 3\text{ A}$$

This simple example could have been solved without the use of superposition, but it illustrates a vital concept for analyzing larger and more complex circuits.

2.5 Thévenin and Norton Equivalent Circuits

*Thévenin*¹³ and *Norton equivalents* are ways to model circuits or circuit sections that may contain various power sources and resistive elements as if they were composed of a single voltage or current source with a series or parallel resistance, respectively.

The key idea was introduced at the very beginning of this chapter: namely, that when we *model* electrical circuits, we choose some level of abstraction that is suitable for the question of interest or the task at hand. Never is a model *identical* to the physical object it is taken to represent. Rather, a useful model allows us to make correct predictions about the behavior of physical objects.

For electrical circuits, these predictions generally concern pairs of current and voltage values that can be simultaneously observed at specific locations. For example, if the voltage across two locations is some value, what current would be measured on a conductor connecting these points?

Before presenting Thévenin and Norton theorems and techniques for determining equivalent circuits, let us introduce the idea of a “one-port” and build some intuition for what it does.

2.5.1 One-ports: Battery and PV Cell

As a general reference, we create an abstraction called a *one-port*.¹⁴ A port consists of two electrically conducting terminals that can have some voltage across them (i.e., potential difference between them), and some current through them, where it is understood that the current entering one terminal must equal the current exiting the other. A one-port as shown in Figure 2.7 is simply a box with a single port. We may or may not know what is inside the box behind the port. An individual element like a resistor could, in principle, be considered a one-port (since it has two ends and something connecting them), but in that case no insight is gained from calling it by that name. One-ports become useful and interesting when there is some active power source inside the box.

A battery or a photovoltaic (PV) cell are concrete examples of one-ports, even though they are in truth quite complicated and nonlinear. While these are physical objects with some internal structure, we focus here on the voltage and current observed at their two terminals. The collection of all physically permissible combinations of voltage and current between two points is called an *operating characteristic*. Because there may be a power source creating this operating characteristic, it can look entirely

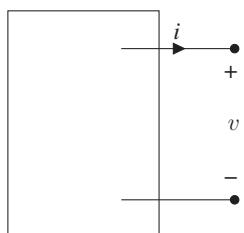


Figure 2.7 A one-port representing a relationship between voltage v and current i , governed by whatever is inside the box.

¹³ In U.S. classrooms, the anglicized pronunciation θévenin is standard, and the accent is usually omitted. In his native French, Léon Charles Thévenin is tévənɛ̃.

¹⁴ One might like to think of the one-port as a “digital twin” of a real physical device.

different than Ohm's law, which applies only to passive devices. In fact, we expect current and voltage from a realistic source to be negatively, not positively correlated.

Rather than trying to faithfully represent the physical devices that actually give rise to this operating characteristic, we create an abstract model from some convenient combination of idealized elements which, if they constituted what's in the box, would produce the same operating characteristic as the one empirically observed. If it is a good model, then its predictions of voltages or currents will match all observations under relevant conditions.

Fundamentally, we will need two kinds of idealized elements: one that gives rise to power (a *source*), and another that somehow restricts or modulates the power flow to make it conform to the operating characteristic. Now, fans of the law of energy conservation will note that we are disregarding the process by which the power or energy comes into the electrical circuit to begin with. In the case of a battery, it is a chemical reaction (see Section 15.3.2); for a PV cell, it is the creation of electron–hole pairs by an incident photon (see Section 14.3.1). But these phenomena are outside our modeling scope. We have two simple choices for how to represent power entering an electrical circuit: a *voltage source*, or a *current source*.

An ideal voltage source is simply an abstraction that says, such-and-such a voltage always exists between these two points. We don't know how that comes to be, and we don't care. We surmise that it will take someone or something doing physical work—especially when there is also a current flowing. In fact, instantaneous power will be the product of current and voltage. If current were to flow in the reverse direction, work would be done *by* (not *on*) the circuit and the source becomes a sink. Other elements in the circuit might restrict such options. But our focus here is not to account for energy or power transfer. The ideal voltage source *per se* only posits that the voltage exists, period. In this chapter, we consider only direct-current circuits and thus constant d.c. voltages, but the concept extends to a.c. voltage sources.

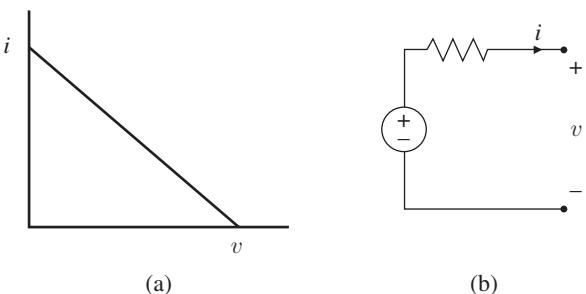
We keep repeating “ideal” because no real-world voltage source will fully conform to that description. Take a AA battery, for example. With an open circuit (no load connected), the AA battery appears as a voltage source: it presents a voltage difference of about 1.5 V between its terminals. We would like to think that this voltage remains constant as some load is connected to the battery and current flows. However, in reality, as an increasing current is drawn, the terminal voltage will decline. This phenomenon can be observed on batteries that have an inexpensive battery level indicator, which is essentially just a voltmeter.¹⁵ For example, an electric bike might indicate a lower battery level while riding uphill, but then recover. The decline in voltage at higher current levels is due to chemical and physical phenomena inside the battery that essentially require each ion to do more work, thus causing losses (which can be observed as the battery heating up). Without going into the details of these phenomena, though, the operating characteristic—at a given state of charge—will simply show the declining terminal voltage at any given current. For some threshold, the short-circuit current, the battery voltage will go to zero.¹⁶

We might observe that the battery voltage tends to decline in a roughly linear way with current. Let us assume here that the current–voltage relationship is exactly linear over the entire range of possible operating conditions. In that case, the actual physical and chemical phenomena can be represented as if they were a simple electrical resistance inside the battery, placed in series with

¹⁵ The open-circuit voltage is a reasonable but inexact proxy for the state of charge (SOC), which depends on the available amounts of chemicals inside the battery. Larger, more expensive batteries have sophisticated algorithms for estimating the SOC. In a lead-acid battery, one can measure the concentration of sulfuric acid—but not without opening up the case.

¹⁶ The reason 12-volt car batteries can cause nasty injuries is that they are designed to sustain a very large short-circuit current before internal resistance draws down their voltage.

Figure 2.8 A simple, linear battery operating characteristic (a) and Thévenin equivalent circuit (b).



an ideal voltage source. This will produce an operating characteristic as in Figure 2.8a that is just a straight line, sloping down from the open-circuit voltage at zero current to a zero voltage at the maximum short-circuit current that the battery can sustain. At this maximum current, the voltage drop across the series resistor will be equal and opposite that of the ideal voltage source, so from the outside we observe zero voltage. The circuit diagram corresponding to this battery model is illustrated on Figure 2.8b. It has the format of a Thévenin equivalent circuit, since it comprises an ideal voltage source and a series resistance.

Let's consider another kind of power source, a PV cell, which has a very different physical mechanism than a battery. Here, it is most intuitive to represent the origin of electrical energy in the circuit as a current source, since photons striking the semiconductor material energize free electrons that then travel as current.

An empirical look at the operating characteristic of a PV cell reveals a nonlinear relationship called the *I–V* curve, shown in Figure 2.9a. (Note that this curve will vary with the operating condition, such as the amount of sunshine.) Readers familiar with semiconductors and electronics might recognize the shape as resembling a *diode* curve, upside down. Without going into the details here, we note that a diode is a device which, owing to its internal physical structure, allows current to pass in one but not the other direction (for more on this, see Chapter 14). Unlike a resistor, though, there is not a constant ratio between current and voltage. Instead, the diode blocks current over a range of voltages, until its conductance increases rapidly at some value (the “knee” of the curve).

If we place a diode in parallel with a current source, oriented such that the conducting direction of the diode faces opposite that of the current source, we get a one-port whose operating characteristic matches that of a PV cell. We might rationalize the picture as follows: The current source is always on. When there is no voltage across the port, the diode is inactive, since no current is driven through it. In that case, the current at the port (also called the short-circuit current, since the port is “shorted out” with no voltage across it) must be equal to the ideal current source. With an increasing positive voltage, the diode current increases. Since, by KCL, this is subtracted from the ideal source current to give the net output current for the port, this net current declines along the curve.¹⁷ The resulting equivalent one-port model for an ideal PV cell is shown on Figure 2.9b.

It's important to reemphasize that an equivalent circuit model is not intended to provide insight into what a device actually looks like inside, or why it behaves the way it does. For instance, a PV cell does not consist of a diode separate from a current source. Rather, these physical phenomena, along with resistance, are completely intertwined within the semiconductor material. The equivalent

¹⁷ The operating characteristic does not show what would happen if a negative voltage were applied. In that case, the diode would be nonconducting, and the current source would have to consume rather than inject power into the circuit. This situation is outside the normal operating specifications, and a PV cell can actually be damaged in this way.

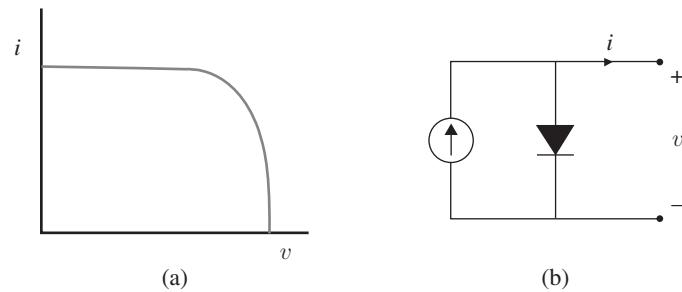


Figure 2.9 (a) The current–voltage operating characteristic or I – V curve for an ideal photovoltaic cell, and (b) the circuit model that reproduces this behavior using an ideal current source and diode. A more realistic model would include a series and parallel resistance for losses within the cell.

circuit simply gives us the permissible set of current–voltage combinations that is consistent with the outward-facing behavior of the PV cell as a one-port.

The equivalent circuit for a PV cell vaguely resembles a Norton equivalent circuit in that it is composed of an ideal current source plus a parallel element that serves to obstruct current in some way. However, in a Norton circuit, the equivalent parallel element must be a resistor, which is linear. But PV cells, and the diodes that help represent them, are nonlinear elements: the I – V curve is not a straight line. Therefore, a PV cell can never be represented by a Norton or Thévenin equivalent circuit.

The reason we chose a simplified battery and PV cell as illustrative examples here is that their representation as voltage and current sources, respectively, is somewhat intuitive and bears some resemblance to their actual physical workings. Yet from the standpoint of a circuit model to reproduce the one-port operating characteristic, we could choose to represent either type of source in either way; there is no requirement that the circuit model reflects any of the internal structure or function. Thus, it is possible to convert the representation of any linear one-port from a Thévenin equivalent to a Norton equivalent circuit, and *vice versa*.

2.5.2 Thévenin and Norton Theorems

Thévenin's theorem states that any black box with two terminals that contains only voltage sources, current sources, and well-behaved resistors (i.e., linear and time-invariant resistors, unlike the diode curve) can be represented as an equivalent one-port containing a single ideal voltage source and a single resistance in series. This is less intuitive in the abstract, but it follows the same principle of reproducing a particular operating characteristic as we followed for the case of the battery.

The constraint that the black box contains only voltage sources, current sources, and resistors—although it may contain any number of them—guarantees that the black box behaves as a *linear* circuit. In other words, the relationship between voltage and current at the terminals of the one-port must be some sort of straight line.

This can be seen as follows: For resistors, the only way that they can all be connected somewhere inside the box is in some aggregate of series and parallel combinations. But any series or parallel combination of resistors is equivalent to some single resistance value, which can be calculated as in Section 2.2.3 on Network Reduction. For multiple voltage and current sources, superposition applies. Since each source in combination with resistance yields a linear operating characteristic, their superposition—the sum of voltages and currents—will be a sum of straight lines, which is itself a straight line. Therefore, connecting only voltage sources, current sources, and resistors in whatever arbitrary way can't give any operating characteristic other than a straight line.

The values of the ideal voltage source and series resistance in a Thévenin equivalent circuit—called the Thévenin voltage, V_{Th} and the Thévenin resistance, R_{Th} —can be determined in two ways: empirical or analytic. If we don't know what's in the black box, we have to make physical measurements. In case the pairs of voltage–current values don't fall on a straight line, whatever is in the box doesn't qualify as a linear circuit. If they do fall on a straight line, we just need to read off the intercept for V_{Th} and the negative slope for R_{Th} . If we do know what's in the box or have a circuit model for it, we can follow a sequence of procedures for network reduction and simplification. A very simple example is shown below.

Norton's theorem is the *dual* or complementary opposite of Thévenin's theorem. It states that any black box with two terminals that has only voltage sources, current sources, and well-behaved resistors inside can be represented as an equivalent one-port containing a single ideal current source with a single resistance in parallel.

Again, the components dictate a linear relationship between current and voltage. The effect of all power sources inside the black box is summed up and expressed in terms of a single current source. While the Thévenin voltage source describes the maximum voltage under an open-circuit condition, which is diminished in actual operation due to a nonzero current, the Norton current source describes the maximum current under a short-circuit condition, which is diminished in actual operation due to a nonzero voltage drop across the one-port.

All the resistive elements inside the box are represented by a single Norton resistance in parallel with the current source. This arrangement can reproduce any linear operating characteristic at the terminals of the one-port, defined by a slope and an intercept—where the intercept is now on the current axis instead of voltage.

We need not worry about circulating currents inside the box. Taken literally, the Norton equivalent circuit implies that while the current source is always working maximally, the parallel circuit branch undermines this effort with a reverse current flow, in an amount depending on the terminal voltage. This might seem wasteful and somehow wrong. But remember that we are not to take Thévenin or Norton equivalent circuits literally: they are models whose purpose is to reproduce the same behavior as that observed from outside the one-port. The only physically real current is the sum of the two branches observed at the terminals.

Since a linear circuit can be represented in either Thévenin or Norton format, these equivalent circuits can be converted into one another. The resistance stays the same, and the values for the respective voltage and current sources, V_{Th} and I_{No} , are converted by Ohm's Law:

$$R_{\text{Th}} = R_{\text{No}}$$

$$V_{\text{Th}} = I_{\text{No}} R_{\text{Th}}$$

$$I_{\text{No}} = \frac{V_{\text{Th}}}{R_{\text{No}}}$$

Example

The one-port in Figure 2.10 has an open-circuit voltage of 18 V, which reduces to 15 V at a load current of 1 A. Assuming a linear operating characteristic, determine the Thévenin and Norton equivalent circuits for this one-port.

The Thévenin voltage is simply the open-circuit voltage, $V_{\text{Th}} = 18 \text{ V}$. The resistance can be deduced from the slope of the operating characteristic: $R_{\text{Th}} = R_{\text{No}} = 3 \text{ V} / 1 \text{ A} = 3 \Omega$. The Norton current is $I_{\text{No}} = V_{\text{Th}} / R_{\text{Th}} = 18 \text{ V} / 3 \Omega = 6 \text{ A}$. Thus, an 18-V voltage source with a 3-Ω resistance in series

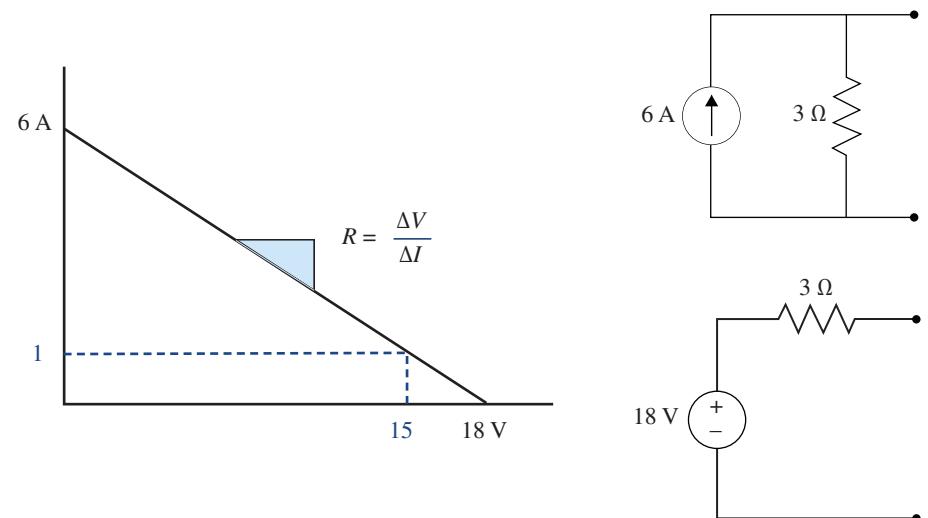


Figure 2.10 Operating characteristic for a linear one-port with Thévenin and Norton equivalents.

behaves in a way that, from the outside, is indistinguishable from a 6-A current source with a 3Ω resistance in parallel.

Example

Convert the circuit illustrated in Figure 2.11 to a Thévenin and Norton equivalent.

When working from a given circuit diagram, there are two separate simplification steps that can be taken in either order: determining the output voltage at the port terminals for a zero-load condition with all sources active (this will be V_{Th}), and determining the network resistance with all sources “deactivated” (this will be $R_{Th} = R_{No}$). Note that while the port terminals are already identified in this figure, in general, the point of simplification for a circuit (i.e., the definition of where it interfaces with the outside world) can be chosen anywhere, and there would be a different Thévenin and Norton equivalent if evaluated from a different vantage point.

For the zero-load condition, it is important to recognize that there will be no current in the circuit branch leading only to the output port (with the 0.5Ω resistor). This means that any resistance on that branch is neglected. In the example, we are left with the 6-A current source and two 1.5Ω resistors in series. In that operating condition, there is a total voltage drop of $V = 3\Omega \cdot 6\text{ A} = 18\text{ V}$

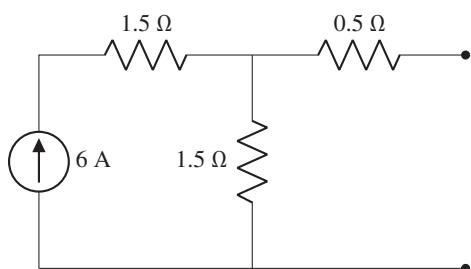


Figure 2.11 Sample linear circuit.

across the two resistors. Since the output terminal connects to the point midway between the two active resistors, it “sees” half the voltage drop, or 9 V. This is the open-circuit voltage V_{th} .

To evaluate the resistive network, we deactivate the current source by setting the current to zero. This amounts to replacing the current source with an open circuit (since there is zero current flowing through an open circuit), or just erasing that circuit branch. In the case of a voltage source, we would set the voltage to zero, which amounts to replacing the source with a short circuit (since there is zero voltage drop across a perfect conductor). This technique extends to more complicated circuits, where all current sources are replaced with open circuits and all voltage sources with short circuits. With the current source in our example at zero, we eliminate the 1.5Ω resistor adjacent to it, and are left with an equivalent circuit loop of a 0.5Ω and a 1.5Ω resistor in series with each other. This gives an equivalent resistance $R_{\text{Th}} = R_{\text{No}} = 2\Omega$.

In the Thévenin representation, the equivalent circuit has a 9-V voltage source in series with a 2Ω resistance. For the Norton representation, we can simply find $I_{\text{No}} = V_{\text{Th}}/R_{\text{Th}} = 4.5\text{ A}$, where the 2Ω resistance is now in parallel. This is shown in Figure 2.12.

If we visualize the one-port in terms of its operating characteristic, we see V_{th} and I_{No} as the voltage and current axis intercepts, and the resistance as the slope. Note that the straight-line operating characteristic is determined by any pair of points on it, or one point and a slope. Our process of evaluating the circuit under a no-load condition amounts to determining one convenient point on the line (namely, the voltage intercept where current is zero), and evaluating the resistive network absent any sources gives the slope.

To convince ourselves that both Thévenin and Norton equivalents correctly reproduce the one-port behavior, we might choose any point on the operating characteristic and check if the values are consistent with these circuits. For instance, consider the point 3 V, 3 A. In the Thévenin circuit, this operating condition would see a voltage drop of 6 V across the 2Ω resistor at 3 A, which properly adds up with the 3-V output to give $V_{\text{Th}} = 9\text{ V}$. In the Norton circuit, the same operating condition would have a current of 1.5 A through the parallel 2Ω resistor, properly adding to the 3-A output current to give $I_{\text{No}} = 4.5\text{ A}$.

The reader might still feel bothered that there must be something not truly equivalent about these circuit models. Indeed, that intuition is correct: where the equivalent circuits fall short is in representing power dissipation *inside* the black box. Since we are defining the one-port only in terms of observed output current and terminal voltage, one question that falls strictly outside our scope of analysis is whether it gets warm inside during operation!

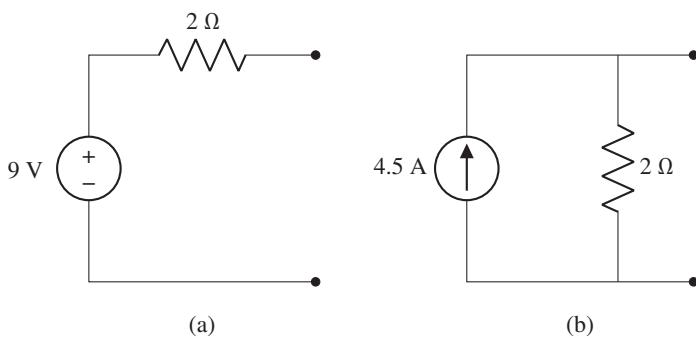


Figure 2.12 Thévenin (a) and Norton (b) equivalents of the circuit in Figure 2.11.

When we apply $P = I^2R$ inside the Thévenin and Norton equivalent circuits in the above example, we obtain different answers for how much power is being delivered by the source and dissipated by the 2Ω resistor in each case (18 W by R_{Th} or 4.5 W by R_{No}). The sources are also doing a different amount of work: At the 3-V , 3-A operating point, the voltage source is delivering 27 W and the current source 13.5 W total. If we chose a different point on the operating characteristic, the current source might be working harder. But all this is beyond scope, since the business of equivalent circuits focuses exclusively on voltage and current *at the port*. Here we observe 3 V and 3 A , meaning 9 W of power delivered to an external load in either case, with no regard for what else might be transpiring inside the black box.

2.6 Magnetic Circuits

In Section 1.5, we introduced the notion of a magnetic field as a pattern of directional forces resulting from the movement of electric charge. We first described fields as analytic artifacts, or maps indicating what would happen to a test object situated in a particular space. We also argued that the field can be appropriately regarded as a physical entity in and of itself, despite the fact that it is devoid of material substance. In the context of magnetic circuits, the latter way of conceptualizing the field is more apt. Here we definitely think of the magnetic field as a “thing” in its own right: something that is present or absent, or present in a particular amount. And, because the presence of a magnetic field indicates that an object can be moved by it—that is, physical work can be done on the object—we consider the field (magnetic or electric) as containing stored or potential energy.

We also stated in Section 1.5.3 that the magnetic field in strict, formal terms represents the density of another quantity, the *magnetic flux*, ϕ . This flux is a measure of something imagined to flow, for example, flowing in circles around a current-carrying wire. A magnet can be represented in terms of a continuous, more or less circular (depending on the magnet’s shape) flow of magnetic flux along the familiar “lines” of the magnetic field. This flux is denser (the lines closer together) inside the magnet and very close to it, and it becomes less dense with increasing distance. In this representation, a solid bar magnet is basically indistinguishable from an electromagnet created by a coil of wire (except for some subtleties around the edges).

The flux representation establishes a crucial property of magnets: namely that they always appear as having two poles, rather than occurring as a single north or south “monopole.” This property can be elegantly expressed in the mathematical statement that the magnetic flux is always continuous, just as if it were a material flowing. Flux lines neither begin nor end anywhere, but perpetually travel around in closed loops. If one enclosed a space with a hypothetical boundary (such as a balloon), no more and no less flux could enter than leave this boundary, since otherwise the flux would have to be created or destroyed within the enclosed space.¹⁸ If a single magnetic north (or south) pole existed, it would violate this condition, since flux would emanate from it (or be absorbed by it) in all directions.

The flux, then, is the “stuff” that must always come back around. In this way, it is analogous to an electric current traveling through a closed circuit, which, as we have seen specifically in Section 2.3.2, also has the property that “what goes in must come out.” This analogy extends in such a way that we can speak of magnetic circuits that obey similar rules as electric circuits. In an electric circuit, the flow of charge can be associated with identifiable material particles; what is flowing

¹⁸ In mathematical notation, this statement constitutes one of *Maxwell’s equations*.

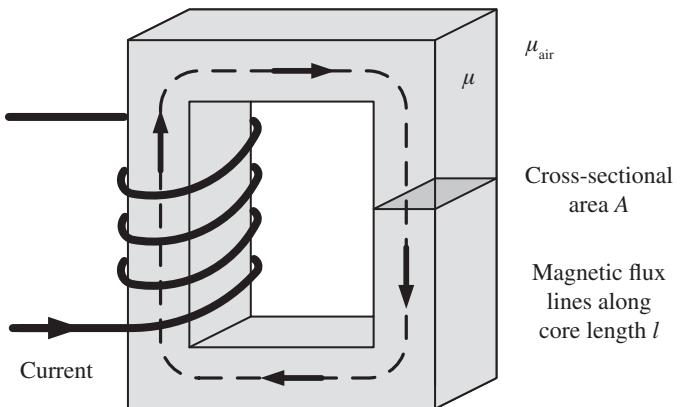


Figure 2.13 A magnetic circuit.

in the magnetic circuit is just the abstract “flux.” Thus, the concept of a magnetic circuit works as an analytic device that keeps track of the quantity of flux and also relates it to the properties of the materials that provide the circuit’s physical path. Figure 2.13 illustrates a magnetic circuit.

Analogous to the electrical resistance, there is a property called *reluctance* that indicates the relative difficulty or ease with which the magnetic flux may traverse an element within a magnetic circuit. Reluctance is denoted by the symbol \mathcal{R} ; it is an *extensive* property that depends on an object’s size and shape. The corresponding *intensive* material property (analogous to electrical conductivity) is the magnetic *permeability*, denoted by μ (Greek lowercase mu). The reluctance of a magnetic circuit element is given by

$$\mathcal{R} = \frac{l}{\mu A}$$

where l is the length and A the cross-sectional area of the magnetic circuit element.

The permeability μ can be regarded as a material’s propensity to carry magnetic flux in response to an externally applied magnetic field. This is like a conductor’s propensity to allow a current to flow through it in response to an applied electric potential difference (electric field; see Section 1.5.2). We can visualize permeability in a vague sense as the ability of the material’s particles to align with an externally applied magnetic field. However, there will be some flux even in the absence of a substantive medium. Thus, the permeability of a vacuum is not zero. Rather, it is assigned a value (so as to keep all other units and measurements consistent) of $\mu_0 = 4\pi \times 10^{-7}$ henries per meter (H/m), and μ_0 is called the *permeability constant*.

The fact that magnetic flux can exist anywhere gives rise to an important practical difference between electric and magnetic circuits: for an electric circuit, because the conductivity of metal is so many orders of magnitude greater than that of the surrounding air, it is a very good approximation to say that all the charge remains confined to the conducting material—such a good approximation, in fact, that we rarely stop to consider it. Magnetic flux, on the other hand, is “messy” to contain because the permeability of air is still some appreciable fraction of that of the best magnetic materials. Thus, the amount of leakage flux, or flux “spilling out” over the edges, is often significant and must be explicitly considered in engineering analyses (e.g., Section 8.5.1).

For an actual material under the influence of a magnetic field, the permeability is not constant, but changes with the field strength or the degree of magnetization. This makes magnetic circuits generally nonlinear. Moreover, the permeability depends on the material’s recent history, that is,

whether it is in the process of being increasingly magnetized or demagnetized.¹⁹ Thus, the value of μ over a range of conditions during the actual operation of magnetic elements must be obtained empirically. Such data or *magnetization curves* are published for the various materials in common use for these purposes, primarily different types of iron and steel.

Quantitatively, we can write the relationship

$$\mathbf{B} = \mu \mathbf{H}$$

where \mathbf{B} is the flux density through the medium (flux per area), which is the same as the familiar magnetic field, and \mathbf{H} is called the magnetic field intensity or *magnetic field strength*.²⁰ \mathbf{B} and \mathbf{H} are written in boldface notation to indicate that they are vector quantities: they have an associated direction, which in this case is the same for both.

As we know, an electric current “chooses” to flow along the path of least resistance. More rigorously speaking, the amount of current flowing through any one circuit branch is inversely proportional to its resistance, that is, directly proportional to its conductance. Similarly, magnetic flux tends toward regions of high magnetic permeability. Thus, while the total amount of flux is constrained by boundary conditions (just like the total amount of current through all circuit branches might be fixed), its density can be distributed throughout space, depending on local variations in magnetic permeability. In this way, the magnetic *field* or *flux density* can be concentrated in a certain region, as if the flux lines were “gathered up.” This effect is used in almost all electric power devices that rely on magnetic fields, by “guiding” the magnetic field through appropriately shaped pieces of iron or steel.

However, unlike an electric current that is confined exclusively to the conducting material of a circuit, the confinement of magnetic flux inside the high-permeability material is always less than perfect. Consequently, an important issue in the design and operation of devices using magnetic flux is the so-called *leakage flux*. This leakage flux is simply the difference between the total amount of flux produced by an electric current and the amount that is successfully confined within the desired region.

The generation of magnetic flux by an electric current can be described in terms of a *magnetomotive force (mmf)* for short), which is analogous to the voltage or electromotive force in an electrical circuit. The *mmf* depends on the amount of current and the configuration of the wire carrying it: specifically, it is not the shape of the wire that matters, but the number of times that the area in question is encircled by this wire in loops or “turns.” Thus,

$$\text{mmf} = Ni$$

where i is the current and N the number of turns.

It is also possible to write an equation for flux analogous to the relationship between voltage and current for a magnetic circuit, where the flux is given by the ratio of *mmf* to the magnetic reluctance of the region within the turns of wire:

$$\phi = \frac{\text{mmf}}{\mathcal{R}}$$

Unlike electric resistance, which for many materials remains constant to a good approximation over a range of voltages and currents (Ohm’s law), the permeability, and thus the reluctance of magnetic materials, varies as the magnetic flux increases and the material becomes “saturated.”

¹⁹ The property of a material to follow a different “path” during magnetization and demagnetization is known as *hysteresis* (see Figure 8.5).

²⁰ These easily confused terms are given here for the sake of completeness. In the power engineering context, phenomena are less often described in terms of fields or field intensity, while the terms flux and permeability are frequently encountered.

In other words, magnetic circuits are not linear. In general, the relationship between *mmf* and flux is not at all straightforward (especially if considered over a wide range of values) and must be determined experimentally. However, the reluctance \mathcal{R} shares with the electric resistance the property that it is additive for elements connected in series, and thus the behavior of an entire magnetic circuit can be derived from its components.

Magnetic flux lines that pass through the enclosed area of a turn of wire are said to *link* this turn. In general, the *flux linkage* of an element is a measure of the extent to which this element is interacting with magnetic flux in its vicinity; it is denoted by λ (Greek lowercase lambda). Flux linkage refers to two symmetrical interactions: (i) the production of magnetic flux from electric current in the element and (ii) the reverse process, where a current through the element is induced by magnetic flux linking it. The single measure of λ applies to both phenomena simultaneously. For a coil of wire, the flux linkage is given approximately by the product of the number of turns and the flux through them:

$$\lambda = N\phi$$

This formula is approximate because it assumes that all the flux lines intersect every turn of wire. In reality, there will be some leakage flux that only interacts with some parts of the coil.

Finally, in anticipation of Section 3.3.1, where we discuss the *inductance* (L) as a crucial property of electric circuit elements, we can state that the flux linkage is also

$$\lambda = Li$$

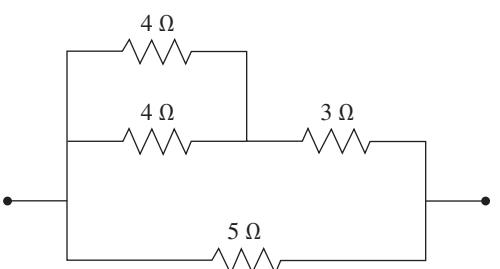
where i is the current through the element. Thus, we can think of inductance as a measure of how much magnetic flux linkage is associated with a given amount of current for a particular device.

Magnetic circuits play a role in electric power systems primarily in the context of generators and transformers, where all the transmitted energy temporarily resides in the form of magnetic fields. Flux lines are used to describe how these devices work, and the analysis of the magnetic circuits they contain, including leakage flux, is crucial in their design. Magnetic flux, and electrical interactions that result from it (specifically, *mutual inductance*), is also important for transmission lines.

Problems and Questions

- 2.1 Calculate the resistance for a parallel combination of two resistors, 5 and 100 Ω . Explain in your own words why it makes sense for the answer to be closer to 5 than to 100.
- 2.2 Find the equivalent resistance of the network in Figure 2.14.

Figure 2.14 Network of resistors.



- 2.3** The resistive network from Figure 2.14 is connected to an ideal 10-V voltage source. Determine the following:
- The voltage across the $3\text{-}\Omega$ resistor.
 - The voltage across either $4\text{-}\Omega$ resistor.
 - The current through the $3\text{-}\Omega$ resistor.
 - The current through either $4\text{-}\Omega$ resistor.
 - The total current delivered by the source.
 - The power delivered by the source.
 - Suppose the $5\text{-}\Omega$ resistor is removed (replaced with an open circuit). Which of the above answers are changed as a consequence?
- 2.4** A string of holiday lights has 50 light bulbs in series. At the end is a receptacle to plug in another string.
- Draw a schematic diagram for two strings with four bulbs each to illustrate the connectivity.
 - If the string plugs into a 120-V outlet, what is the voltage across each bulb?
 - When an individual LED lamp stops working, it usually becomes a short circuit, whereas a failing incandescent light becomes an open circuit. Incandescent holiday lights usually have a shunt wire to turn a bulb into a short circuit when the filament burns out. Explain why this is practical, and explain the effect on the other lights in the string when an individual bulb (incandescent or LED) fails.
 - Somewhere on the box, a warning tells you not to connect more than some maximum number of strings together. Why? Explain what might happen if you connect too many strings.
- 2.5** The following (purely resistive) appliances are found in a studio apartment: a 600-W toaster, three incandescent lights at 100 W each, and an 800-W space heater. Assume a nominal voltage of 120 V.
- What is the resistance of each load?
 - Suppose the five loads are plugged into a power strip and into the same wall outlet, which is rated for a maximum current of 15 A. Will this 15-A circuit be able to support all these loads operating at the same time?
 - Suppose the power strip with all five loads is plugged into the wall outlet through a long extension cord. The cord has a resistance of $0.2\ \Omega$ (counting both conductors). How much power is dissipated in the cord? Do you think this might be a problem?
- 2.6** A power distribution line serves a load of 0.50 MW, 10 km from the source. (Ignoring the more complicated aspects of power lines such as reactive power and multiple phases, we assume a power factor of unity and a single phase, meaning two conductors.) Search online to find product specifications for “Raven” and “Quail” ACSR (Aluminum Core Steel Reinforced) conductors. State answers to two significant figures.
- If the distribution line is operated at 4.0 kV (the difference between the two conductors, as measured at the load end), what is the line current?
 - If the distribution line is operated at 12 kV instead, what is the line current?
 - What is the resistance of the 10-km line using 1/0 (pronounced “one aught”) ACSR conductor “Raven”?
 - What is the resistance of the 10-km line if 2/0 ACSR (“Quail”) is used instead?

- (e) If the line is operated at 4.0 kV and 1/0 ACSR is used, what are the line losses in kW, and expressed as a percentage of the load served?
- (f) What are the losses (in kW and in %) at 4.0 kV, if 2/0 ACSR is used instead?
- (g) What are the losses (in kW and in %) for 1/0 ACSR if the circuit is operated at 12 kV instead?
- (h) What general insight about the design of electric power systems is this exercise intended to convey?
- 2.7** A mysterious box with two metal terminals exhibits the following properties: The voltage between the two terminals, when no current is flowing into or out of the box, is 12 V. When a 24Ω resistance is connected externally between the two terminals, the measured voltage is reduced to 9 V.
- Determine the Thévenin and Norton equivalent circuits.
 - For what external resistance would you expect the terminal voltage to drop to 3 V?
 - How much power in watts is delivered to the external resistance in each operating condition, 9 V and 3 V?
 - Can you determine how much power is actually dissipated inside the box under each operating condition? Explain.
- 2.8** Suppose you've dropped a small metal object in a space you can't reach, and you want to retrieve it. You cleverly think to wrap some coated wire around a large nail and connect the two ends of the wire to a AA battery, so as to create an electromagnet that will pick up the object.
- What do you think is the most likely way you might get injured from this experiment? Explain.
 - Why does it matter if the wire is coated with insulating material?
 - What is the significance of the magnetic permeability μ of the nail and the dropped object?
 - Suppose you have a wide selection of wires and nails in your home workshop. Explain your design choices to make your electromagnet as strong as it can be, without hurting yourself.

3

AC Power

3.1 Alternating Current and Voltage

Many interesting technical characteristics of power systems result from their use of alternating current (a.c. or AC) instead of direct current. In a d.c. circuit, the polarity always remains the same: the potential always stays positive on one side and negative on the other, and the current always flows in the same direction. In an a.c. circuit, this polarity reverses and oscillates very rapidly. For power systems in the United States, the *a.c. frequency* is 60 hertz (Hz) or 60 cycles per second, meaning that the direction of voltage and current are reversed, and reversed back again, 60 times every second.

3.1.1 Historical Notes

The main reason for using a.c. in power systems is that it allows raising and lowering the voltage by means of transformers. As we will see in Chapter 8, transformers cannot be operated with d.c. Today it is quite easy to change the voltage in a d.c. circuit, but it requires far more sophisticated equipment (i.e., power electronics) that had not been invented yet in the early days of electric power. The first power systems, which operated on d.c., were therefore limited to rather low transmission voltages: although the generators could have been designed to produce power at a higher voltage, safety considerations at the customer end dictated that the voltage be kept low. Consequently, line losses were a major problem and in effect limited the geographic expansion of power systems. After the transformer was introduced in the 1880s, d.c. and a.c. systems spent some years in fierce competition (the “Battle of the Currents”), with Edison and Westinghouse as prominent advocates on either side. The major obstacles in the way of the alternating current approach—namely, concerns about the safety of high-voltage transmission, as well as the challenge of designing an a.c. motor—were largely resolved by the mid-1890s.¹

The choice of frequency for a.c. power represented a compromise among the needs of different types of equipment. During the early years of a.c. systems, numerous different frequencies ranging from 25 to $133\frac{1}{3}$ cycles were used.²

For rotating generators, lower a.c. frequencies are somewhat advantageous in that they require fewer magnetic poles inside the rotor to step the mechanical speed up to the electrical speed

¹ For a thorough and fascinating historical discussion of the development of electric power systems, see Thomas P. Hughes, *Networks of Power: Electrification in Western Society, 1880–1930* (Baltimore, MD: Johns Hopkins University Press, 1983).

² See Hughes, Networks of Power, pp. 127ff., and Benjamin G. Lamme, The Technical Story of the Frequencies, *IEEE Transactions* 37, 1918.

(see Section 10.3.2). This constraint became less significant as high-speed steam turbines supplemented and replaced slow-moving hydroturbines and reciprocating steam engines. On the other hand, machines and transformers can be made smaller at higher frequencies, because the induced a.c. electromotive force depends on the rate of change of magnetic flux (see Section 10.1). At a faster rate of alternation, a smaller amplitude of magnetic field is required, allowing for smaller magnetizing currents and magnetic cores.

For a.c. transmission, lower frequencies are desirable because a line's *inductive reactance* increases with frequency and constrains the amount of power that can be transmitted on a given line (see Sections 3.3 and 7.3). For loads, higher frequencies have historically been preferable—particularly for incandescent lamps, which flicker with dissipated power at twice the a.c. frequency. This flickering becomes more and more noticeable to the human eye at lower frequencies. Today, an increasing fraction of loads are power electronic, and thus rather indifferent to a.c. frequency.

Historically, the above factors had led to the use of different frequencies in various local a.c. systems, but at some point interoperability began to outweigh optimization for specific components. After consideration of the different types of equipment already in use and the prospects for adapting new designs, efforts to standardize power frequency finally resulted in convergence to a 60 cycle standard in the United States and 50 cycles in Europe. North, Central and northern parts of South America along with a few other countries today use 60 Hz, while Africa, Australia, and a majority of Asian countries use 50 Hz.

There remain some intriguing cases of different a.c. frequency choices in niche applications. One of these is the network of electric railroads in Europe, served by its own single-phase, long-distance a.c. transmission system operated at 16.7 Hz. This frequency works fine for the electric train locomotives, and it is more efficient to transmit than 50 Hz due to the lower line reactance. Another case is aircraft a.c. power, which has a standard frequency at 400 Hz. On the scale of an airplane, conductor impedance is negligible, but weight matters a lot. Before the advent of solid state power converters, generators and transformers onboard aircraft could be made smaller and lighter at the higher frequency.

3.1.2 Mathematical Description of Alternating Current

A sine wave represents the cyclical increase and decrease of a quantity over time. The oscillation of voltage and current in an a.c. system is (approximately) modeled by a sinusoidal curve, meaning that it is mathematically described by the trigonometric functions of sine or cosine.³ In these functions, time appears not in the accustomed units of seconds or minutes, but in terms of an *angle* that can be expressed in units of degrees or radians.

A sinusoidal function as illustrated in Figure 3.1 is specified by three parameters: *amplitude*, *frequency*, and *phase*. The amplitude gives the maximum value or height of the curve, as measured from the neutral position. (The total distance from crest to trough is twice the amplitude.) The frequency gives the number of complete oscillations per unit time. Alternatively, one can specify the rate of oscillation in terms of the inverse of frequency, the *period*. The period is simply the duration of one complete cycle. The phase indicates the starting point, or offset in time compared

³ Historically, with power generated from rotating machines (Section 10.1), this approximation is so good that departures from the sinusoidal model rarely had to be considered in power engineering practice, and would rarely be part of an introductory curriculum. We address the imperfect representation of the a.c. signal in the context of phasors (Section 3.5) and phasor measurement (Section 16.2.4), and turn explicitly to distorted waveforms in Section 5.3.

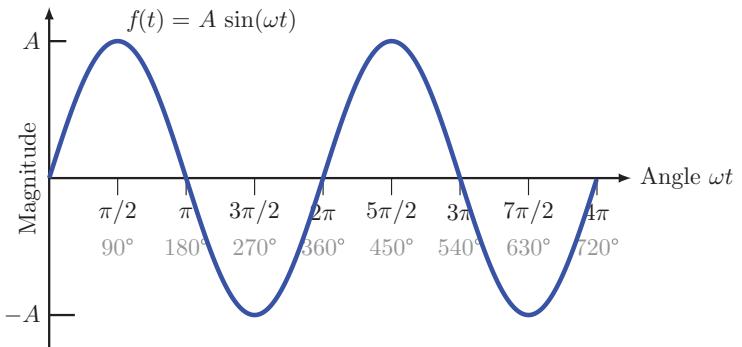


Figure 3.1 A sine function $f(t) = A \sin \omega t$ plotted against time t or, equivalently, angle ωt .

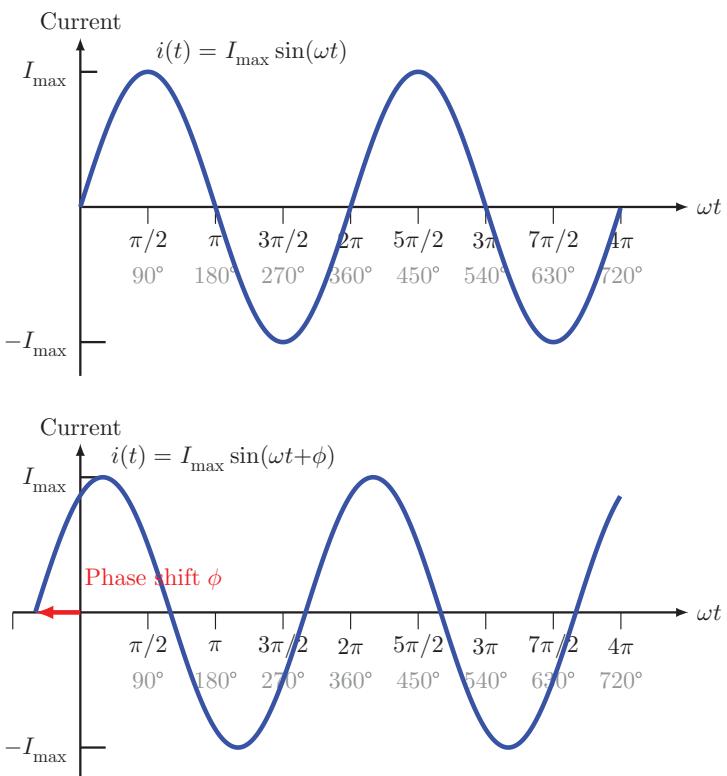


Figure 3.2 Sinusoidal alternating current with phase shift.

to the same curve starting at zero. Graphically, we see the phase simply as a shift of the entire curve to the left (positive) or right (negative phase angle). The phase angle is often denoted by ϕ (Greek lowercase phi). For example, the lower curve in Figure 3.2 appears shifted to the left, which corresponds to an advancement (positive ϕ) compared to the upper curve.

Expressing time as an angle allows us to take a sine or cosine of that number. For example, the sine of 30° is $\frac{1}{2}$, but there is no such thing as the sine of 30 seconds. The argument of a sinusoidal function (the variable or object of which we find the sine) must be dimensionless—that is, without

physical dimension like time, distance, mass, or charge. An angle, though measured in units of degrees or radians, has no physical dimension; it really represents a ratio or fraction of a whole. This is consistent with the fact that a sine function represents a relationship between two quantities: in a right triangle, the sine of one angle is the ratio of lengths of two of the sides (the reader may recall “opposite over hypotenuse” for sine, and “adjacent over hypotenuse” for cosine). Time as an angle means time as some fraction of a whole.

Turning time into an angle or fraction is logical only because in an oscillation, time is cyclical; the process repeats itself. We do not care whether we are on our first or four-hundredth swing, but rather about *where* in the oscillation we find ourselves at a given instant. One complete oscillation, the duration or period of which would be 1/60th of a second for 60 Hz, is taken to correspond to a full circle of 360°. Any angle can be understood, then, as specifying a fraction or multiple of that complete oscillation. Knowing the period of the oscillation from context, we may interpret degrees or radians as representing amounts of time.

Plotted against angle on the horizontal axis, the height of the sine curve is simply the value of the sine for each angle, scaled up by a factor corresponding to the amplitude. As the angle is increased, it eventually describes a complete circle, and the function repeats itself.

In the context of sinusoidal functions, angles are often specified in units of *radians* (rad) rather than degrees. Radians refer to the arc described by an angle. The conversion is simple. Since the circumference of an entire circle is given by $2\pi r$, where r is the radius and π (pi) = 3.1415..., 2π radians correspond to 360°. (The radius is left out since the size of the circle is arbitrary.) Any fraction of a radian, then, represents a fraction of a circle, or number of degrees: π rad = 180° or one-half cycle; $\frac{\pi}{2}$ rad = 90° or one-quarter cycle, and so on. Figure 3.1 illustrates a sine wave with both units of angle.

The frequency of a sinusoidal function is often given in terms of radians per second, in which case it is called an *angular frequency*. Angular frequencies are usually denoted by ω , the Greek lowercase omega, as opposed to f or v (Greek lowercase nu) for frequency in cycles per second or Hertz (Hz). The angular frequency corresponding to 60 Hz is

$$\omega = 2\pi \frac{\text{rad}}{\text{cycle}} \cdot 60 \frac{\text{cycle}}{\text{sec}} = 377 \frac{\text{rad}}{\text{sec}}$$

For 50 Hz, the angular frequency is 314 $\frac{\text{rad}}{\text{sec}}$.

An alternating voltage as a function of time can be written as the following sinusoidal function:

$$v(t) = V_{\max} \cos(\omega t + \phi) \quad (3.1)$$

The quantity V_{\max} is the maximum value or amplitude of the voltage. Since the value of a sine or cosine varies between -1 and +1, the actual voltage oscillates between $-V_{\max}$ and $+V_{\max}$. When we plug in time t in units of seconds, multiplying it by the angular frequency ω gives an argument ωt in radians.

In Figure 3.1 we see no phase shift, as the curve starts with the instantaneous voltage zero at time zero. For the general case, where the voltage does not happen to be zero at time $t = 0$, a phase shift ϕ (Greek lowercase phi, pronounced either “fee” or “fie”) is added. Any sinusoidal function can be just as correctly written as either a sine or a cosine. For example, the curve in Figure 3.1 can be represented either as a sine with $\phi = 0$, or as a cosine $v(t) = V_{\max} \cos(\omega t - \frac{\pi}{2})$. In power engineering, the cosine is preferred by convention, and it is important to keep the format consistent across different functions in case we wish to compare phase angles.

The phase angle ϕ , when used in such an argument, should be in radians to match ωt . In other contexts, the phase angle may be stated in units of degrees: especially in phasor notation

(Section 3.5), where the frequency ω is implicit and there is no need to evaluate $(\omega t + \phi)$. It is always a good idea to check units of angle for consistency.

The phase affects neither frequency nor amplitude of the curve; it simply amounts to a difference in what time is considered “zero.” To distinguish phase shifts between different quantities, we may include a subscript such as ϕ_V for the voltage phase angle.

Sinusoidally varying quantities in a.c. power systems include voltages and current. We make the key assumption that current, like voltage, behaves in a perfectly sinusoidal manner, with the same frequency.⁴ This allows us to write $i(t) = I_{\max} \cos(\omega t + \phi_I)$. In Section 3.3 we will see that the difference in phase angle between voltage and current, $\phi_V - \phi_I$, has physical significance.

3.1.3 The rms Value

For many practical applications, we are only interested in an overall magnitude of these sinusoidal functions, not the details of their oscillation in time. For example, we might care how hot a wire will get if it conducts a certain current. Conceivably, we could just indicate the amplitude of the sine wave, but this would not represent the quantity very well: most of the time, the actual value of the function is much less than the maximum. Alternatively, we could take a simple arithmetic average, or mean. However, since a sine wave is positive half the time and negative the other half, we would just get zero, regardless of the amplitude, so this type of average would contain no useful information.

What we would like is some way of averaging the curve that offers a good representation of how much current or voltage is actually being supplied: a meaningful physical measure; something of an equivalent to a d.c. value. Specifically, we would like average values of current and voltage that yield the correct amount of power when multiplied (see Section 3.4). Fortunately, such an average is readily computed: it is called the *root-mean-square* (*rms*) value.

The rms value is derived by first squaring the entire function, then taking the average (mean), and finally taking the square root of this mean. Mathematically, the rms value is defined for any set of real values or continuous-time waveforms, as the square root of the arithmetic mean of the squares of the values:

$$V_{\text{rms}} = \sqrt{\frac{v_1^2 + v_2^2 + v_3^2 + \dots + v_n^2}{n}}$$

We focus here on the special case of a sinusoidal function, illustrated in Figure 3.3 with curves labeled $v(t)$ and $v^2(t)$, where we take the amplitude $V_{\max} = 1$ for convenience. The squared sine retains the same basic shape, but is compressed in half. Because the squared curve resides entirely in the positive region, we can now take a meaningful average. In fact, because the curve is still perfectly symmetric, its average is simply $\frac{1}{2}V_{\max}$. The remaining (and perhaps slightly counterintuitive) step consists of re-normalizing this average value to the original curve, by taking the square root: basically, we are just going backwards and undoing the step that made the curve manageable for averaging purposes.

Since $1/2$ is less than 1, its square root is greater than itself; it comes to $\frac{1}{\sqrt{2}} \approx 0.707$. The rms value of a sinusoidal function is therefore 0.707 times the original amplitude. This holds true for the general case when the amplitude is not 1.

⁴ This assumption is plausible if we expect circuit elements to obey Ohm’s law, $V = IR$. It does not apply to nonlinear elements such as those appearing in Section 5.3.1.

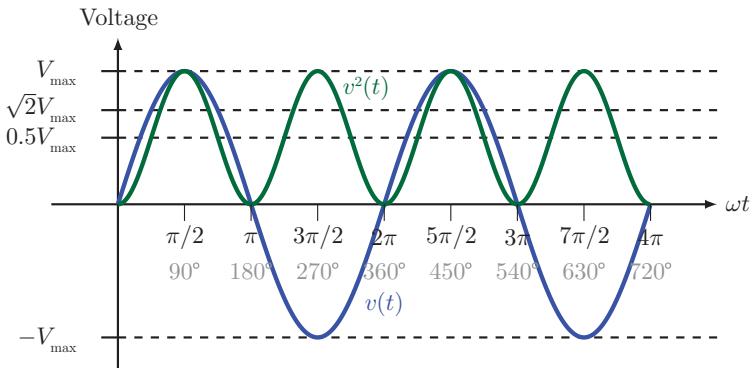


Figure 3.3 Derivation of the rms value.

Utility voltages and currents are almost always given as rms values. For example, 120 V is the nominal rms voltage for a residential outlet in the United States. The rms value is also what a conventional a.c. current or voltmeter will report as a measurement. Note that when the rms voltage and current are multiplied together, the product conveniently gives the correct amount of power transmitted (see Section 3.4).

Example

If 120 V is the rms value of your household voltage, what is the maximum instantaneous or peak voltage?

Since $V_{\text{rms}} = \frac{1}{\sqrt{2}} \cdot V_{\text{max}}$, the maximum instantaneous voltage is $V_{\text{max}} = 120 \text{ V} \cdot \sqrt{2} = \pm 169.7 \text{ V}$.

The maximum instantaneous value of the voltage is of practical interest because it determines the requirements for electrical insulation on wires and other energized parts. In fact, one argument against a.c. in the early days was that it would be less economical than d.c. due to the insulation requirements being effectively twice as high, for the same amount of power transmitted. For current, the instantaneous maximum is usually uninteresting, because current limitations are related to resistive heating which happens cumulatively over time.

3.2 Power for the Resistive Case

Power is a measure of energy per unit time. Power therefore gives the rate of energy consumption, production, or transfer. The units for power are generally watts (W). For example, the watt rating of an appliance gives the rate at which it uses energy. The total amount of energy consumed by this appliance is the wattage multiplied by the amount of time during which it was used; this energy can be expressed in units of watt-hours (or, more commonly, kilowatt-hours).

As we saw for the d.c. case in Section 1.3, the power dissipated by a circuit element—whether an actual appliance or simply a wire—is given by the product of its resistance and the square of the current through it: $P = I^2R$. The term “dissipated” indicates that the electric energy is being converted to heat. This heat may be part of the appliance’s intended function (as in any electric heating device), or it may be considered a loss (as in the resistive heating of transmission lines); the physical process is the same. Note that dissipation implies thermodynamic irreversibility: you will never see ambient heat spontaneously coalescing in a wire to move electrons.

A more general way of calculating power is as the product of current and voltage: $P = IV$. For a resistive element,⁵ we can write Ohm's law as $V = IR$ and substitute it for V to see that the two formulas amount to the same thing:

$$P = IV = I \cdot (IR) = I^2R \quad (3.2)$$

Example

Consider an old-fashioned incandescent light bulb, rated at 60 W. This means that the filament dissipates energy at the rate of 60 W when presented with a given voltage, which is assumed to be standard household voltage (120 V). By convention, rms values are used for both current and voltage, which reflect an average over time. The power dissipated equals the rms voltage applied to the light bulb, times the rms current through it. The current must be half an ampere, since $60\text{ W} = 0.5\text{ A} \cdot 120\text{ V}$.

While the power rating is based on an assumed voltage, the fixed physical attribute of the light bulb itself is its resistance—in this case, 240Ω . We can infer the resistance from Ohm's law, $120\text{ V} = 0.5\text{ A} \cdot 240\Omega$, and verify that the power agrees with I^2R : $60\text{ W} = (0.5\text{ A})^2 \cdot 240\Omega$.

When the household voltage varies, the same light bulb will produce different amounts of power. Why, then, aren't light bulbs labeled with their resistance? Because the power consumption is more informative; voltage doesn't tend to vary too much; and consumers would get confused (if I want a light bulb that makes more light, should I buy the one with a bigger or smaller number of omegas?).

If the voltage is fixed and the current is determined by the resistance, then cutting the resistance in half will double the current, according to $V = IR$. The power $P = I^2R$ will also double, since the current is squared but multiplied again by resistance. So a 120Ω light bulb is twice as bright⁶ (drawing 1 A of current and 120 W of power) as the 240Ω bulb.

3.2.1 Power Dissipated Versus Transmitted

Let us now consider a transmission line, and distinguish between the power *dissipated* and the power *transmitted* by the line. The dissipated power is simply given by $P = I^2R$. We could also write this as $P = IV$, but that would be less convenient, for two reasons.

First, although it is tempting to think of a power line as just a resistive wire, it actually has a significant reactance (introduced in Section 3.3). This will complicate the multiplication of current and voltage, whereas taking the square of the current magnitude is always straightforward.

Second, if we tried to calculate power dissipated on the line by using $P = IV$, we would have to be very careful about which V to use. Remembering that Ohm's law refers to the voltage drop across a resistor—that is, the potential difference between the two ends of a single object—we recognize that the relevant V must be the voltage difference between the two ends of the line, known as the voltage *drop* or “line drop” (see Section 1.3.3), sometimes helpfully labeled as V_d .⁷ This line drop V_d is distinct from the nominal “line voltage,” which specifies the voltage difference either between line and ground, or between one conductor and another.⁸ Typically, the line drop is a few percent

⁵ For the general case where there may also be reactance, see Section 3.3. It will always be true, though, that I^2R gives the power dissipated by resistive heating, if not the entire power transferred to the complex load.

⁶ Not to the human eye, whose light sensitivity is more logarithmic than linear, but it dissipates twice the power.

⁷ If an ideal transmission line had zero resistance, there would be zero voltage drop and no power dissipated by the line.

⁸ The reader may notice that, if there is some line drop, then the line voltage relative to ground must be different depending on *where* it is measured—say, at the beginning or the end of the line. When referring generally to the line voltage, we make an implicit assumption that it describes either an approximate average or a nominal reference value, neither of which would be presumed identical to what is measured in the field at a specific location.

In situations where the difference matters (as in Chapter 9), we explicitly distinguish the voltage at the sending and receiving end of a transmission line.

of the line voltage, but in practice is not known exactly, since it varies with the loading condition in real-time.

For these reasons, thermal losses are better calculated using $P = I^2R$, and are often referred to as “I-squared-R losses.” Note that when calculating line losses for a complete circuit, we have to consider each individual conductor (e.g., two conductors for a single-phase circuit).

If instead we care about the power transmitted to a load, we should think of the transmission line as extended terminals, like battery terminals. (Alternatively, we can visualize a single conductor and consider the ground as the other terminal.) The power that is available to a load connected to this line can be calculated with the formula $P = IV$, but V now refers to the nominal line voltage, which is the difference between the two terminals (or between the single line and ground). We say that the power has been transmitted by the line “at the voltage V .”

$P = IV$ makes intuitive sense when we remember that voltage is a measure of energy per unit charge, while the current is the flow rate of charge. The product of voltage and current therefore tells us how many electrons are passing through, multiplied by the amount of energy each electron “carries.” Energy is carried in the sense that an electron that has been propelled to a higher voltage level has the potential to do more work as it returns to ground. In terms of the units, we can see that the charge cancels, and we are left with the proper units of power:

$$\text{Power} = \frac{\text{Energy}}{\text{Time}} = \frac{\text{Charge}}{\text{Time}} \cdot \frac{\text{Energy}}{\text{Charge}} \quad (3.3)$$

3.2.2 Time-Varying Resistive Power

Let us now take a closer look at a resistor presented with a varying voltage. We can always correctly interpret the statement $P = IV$ when I and V vary in time as a statement of instantaneous conditions. Namely, instantaneous power equals instantaneous current times instantaneous voltage, at any given moment. To be perfectly clear, we should write each quantity as an explicit function of time:

$$p(t) = i(t) v(t) \quad (3.4)$$

Equation (3.4) for instantaneous power *always* holds true, for any moment in time, regardless of the complexities introduced later. However, instantaneous power is often not very interesting or useful in practice. In power systems, we generally care about power transmitted or consumed on a time scale much greater than 1/60 of a second. Therefore, we need an expression for power as averaged over entire cycles of alternating current and voltage.

Fortunately, we already know how to do this. The important realization is that Ohm’s law $V = IR$ also holds for every instant in time. Furthermore, we assume our resistor is well-behaved or *linear*, with a constant resistance.⁹ Consequently, if the applied voltage is a sinusoid, the current through the resistor will be a similar sinusoid, scaled up or down proportionally depending on the value of the resistance. The two curves $v(t)$ and $i(t)$ will only differ in their amplitudes.

To calculate a useful expression for average power, we multiply current and voltage for all times throughout the cycle, and then take the average. The subscript r is a reminder that this equation only holds for a resistive load:

$$p_r(t) = v(t)i(t) = V_{\max} \cos(\omega t)I_{\max} \cos(\omega t) = V_{\max} I_{\max} \cos^2(\omega t) \quad (3.5)$$

⁹ This is almost always an excellent assumption for physical resistors, as opposed to power electronic loads. Even though resistance can vary (for example, with temperature), it will do so much more slowly than on the time scale of a.c. cycles.

It is convenient to reformat Eq. (3.5) by applying the trigonometric identity $\cos^2 u = (1 + \cos 2u)/2$, to write:

$$p_r(t) = V_{\max} I_{\max} \cos^2(\omega t) = \frac{1}{2} V_{\max} I_{\max} + \frac{1}{2} V_{\max} I_{\max} \cos(2\omega t) \quad (3.6)$$

This expression in Eq. (3.6) is a sinusoidal function of twice the original frequency, centered on one-half the amplitude (just as seen in the squared curve of Figure 3.3). With the cosine varying between ± 1 , the value of $p(t)$ varies between 0 and $V_{\max} I_{\max}$. Therefore the average power is just one-half the product of the voltage and current amplitudes:

$$P_{\text{ave},r} = \frac{1}{2} V_{\max} I_{\max} = V_{\text{rms}} I_{\text{rms}} \quad (3.7)$$

Conveniently, this amounts to the same as the product of the rms values, just as they are reported by a.c. voltmeters and ammeters. Working with this expression for power is no different than analyzing d.c. circuits. However, as soon as we expand our scope from pure resistors to general elements of a.c. circuits, we will encounter voltages and currents that are not aligned in time, which will complicate the process of multiplying them together.

3.3 Impedance

In Chapter 1, we discussed electrical resistance as the property of a material or electric device to resist the flow of direct current through it. *Impedance* is the more general term, which accounts for resistance as well as *reactance*.

Reactance is the property of a device to influence the relative timing of an alternating voltage and current. By doing so, it presents a sort of impediment of its own to the flow of alternating current, depending on the frequency. Reactance results from the interplay of electric or magnetic fields and is therefore related to the geometry of a device. It is a physical phenomenon separate from resistance.

There are two types of reactance: inductive reactance, which is based on inductance, and capacitive reactance, based on capacitance. Impedance is the general term that describes any combination of resistance and reactance. Inductance and capacitance are physical properties of objects that depend on their shape (which, for our purposes, is almost always assumed to be fixed). Reactance also takes the a.c. frequency into account; it describes how an inductance or capacitance affects a circuit. Resistance, reactance, and impedance are all measured in ohms (Ω).

3.3.1 Inductance

The classic inductive device is a coil of wire, called an inductor or a *solenoid*. Its behavior results from the physical fact that an electric current produces a magnetic field around it. This magnetic field describes a circular pattern around a current-carrying wire; the conventional direction of the field can be specified with a “right-hand rule.”¹⁰ When a wire is coiled up as shown in Figure 3.4, it effectively amplifies this magnetic field, because the contributions from the individual loops add together. The sum of these contributions is especially great in the center, pointing along the central axis of the coil. The resulting field can be further amplified by inserting a material of high magnetic permeability (such as iron) into the coil; this is how an electromagnet is made.

¹⁰ If the direction of one’s right thumb corresponds to the current flow (positive to negative potential), the curled fingers of that hand indicate the direction of the magnetic field (see Figure 1.2).

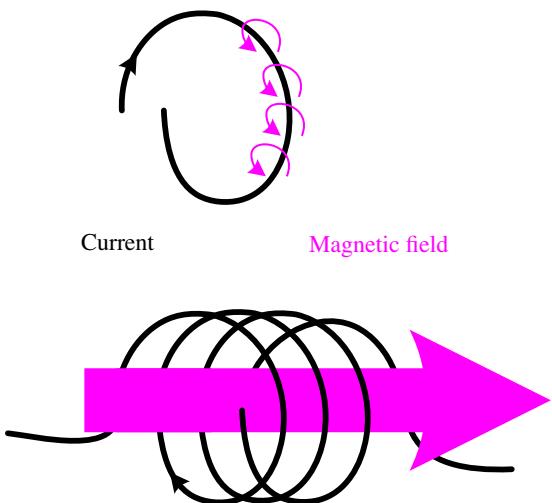


Figure 3.4 A basic inductor, or solenoid.

When such a coil is placed in an a.c. circuit, a second physical fact comes into play: namely, that a *changing* magnetic field in the vicinity of a conducting wire creates an electromotive force, which in turn induces an electric current to flow through this same wire. If the current through the coil oscillates back and forth, then so does the magnetic field in its center—or the magnetic flux linking the coil. Because this magnetic field or flux is continuously changing, it induces another current in the coil, which is proportional to the rate of change of the field.¹¹ The direction of the induced current will be such as to *oppose* the change in the current responsible for producing the magnetic field in the first place. In other words, the inductor exerts an inhibitive effect on a change in current flow.¹²

This inhibitive effect results in a delay or phase shift of the alternating current with respect to the alternating voltage that we assume was externally applied to produce any current to begin with. Specifically, an ideal inductor (with no resistance at all) will cause the current to lag behind the voltage by a quarter cycle, or 90° , as shown in Figure 3.5.

This result is difficult to explain intuitively. We won't attempt to walk through the evolving voltage, current and magnetic field over the course of a cycle. But the graph clearly shows that the current reaches its maximum at the instant that the voltage decreases most rapidly, and that the current increases most rapidly just as the voltage reaches its maximum. This behavior comes from the constitutive relationship between voltage and current for an inductor given in Eq. (3.8), which introduces a time rate of change.

When describing the behavior of electrical devices in the context of circuit analysis, we are generally interested in writing down a mathematical relationship between the current passing through and the voltage drop across the device. For a resistor, this is simply Ohm's law, $V = IR$, where the resistance R is the proportionality constant between voltage and current. It turns out that the inductance L also works as a proportionality constant between current and voltage across

¹¹ The formal definition of the electromotive force, as presented below and in Chapter 10, is in terms of the rate of change of magnetic flux—which could be due to either the strength of the field, or a changing orientation of the field relative to the coil. When nothing is moving in a single inductor, it's not important to distinguish between field and flux linkage.

¹² This result can be derived through right-hand rules, or simply by considering the law of energy conservation: if the induced current were in the same direction as the original (increasing) current, it would amplify the magnetic field that produced it, which would in turn increase the induced current, and so on, indefinitely—clearly an impossible scenario!

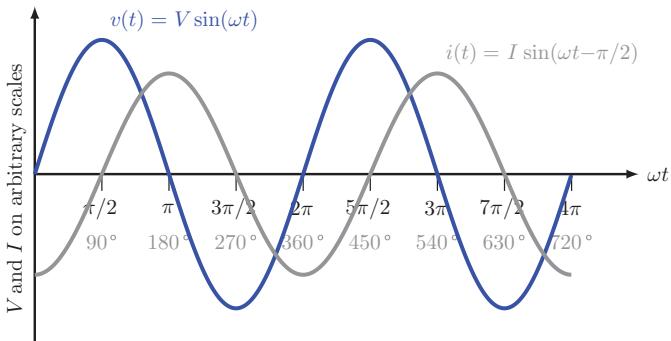


Figure 3.5 Current lagging voltage by 90°. These functions could be written as $V(t) = V_{\max} \sin(\omega t)$ and $i(t) = I_{\max} \sin(\omega t - \pi/2)$. Note that the relative amplitudes are unimportant, because they are measured in different units.

an inductor, but in this case the equation involves the *rate of change* of current, rather than the actual value of current at any given time. Inductance is measured in units of henrys (H).¹³

Readers familiar with calculus will recognize the notation $\frac{di}{dt}$, which represents the time derivative, or rate of change (of current, in this case) with respect to time. Thus, we write:

$$v(t) = -L \frac{di}{dt} \quad (3.8)$$

Equation (3.8) states that the voltage drop v across an inductor at any instant must equal the product of its inductance L , and the rate of change of the current i through it. The negative sign indicates that the voltage and the change in current are in opposite directions, as noted above: it is an inhibitive, not an amplifying effect. The behavior defined in Eq. (3.8) constitutes what we call an “inductor.” This equation can be used in circuit analysis in a manner analogous to Ohm’s law to establish relationships between current and voltage at different points in the circuit. The time derivative just makes it more cumbersome to manipulate.

Casually speaking, we might say that this equation characterizes an inductor as the kind of thing that does not like to see current change: it resists, or “reacts to,” any change in current. This is consistent with the idea that any change in current will cause a changing magnetic field that acts to oppose the new current. The greater the inductance, the more voltage is required to effect a change in current. Conversely, if the current is forced to change dramatically—say, by being interrupted elsewhere in the circuit—we will observe a voltage spike caused by the inductor’s reaction to this change.

The constitutive Eq. (3.8) can also be written with an \mathcal{E} for *electromotive force*, rather than with a V or $v(t)$. This derives from Faraday’s law of induction,

$$\mathcal{E} = -\frac{d\Phi_B}{dt} \quad (3.9)$$

where Φ_B is the magnetic flux given by

$$\Phi_B = L i(t) \quad (3.10)$$

If we now take the time derivative of Eq. (3.10), we get Eq. (3.8) in the form

$$\mathcal{E} = L \frac{di}{dt} \quad (3.11)$$

¹³ Equivalently, 1 H = 1 V·s/A = 1 Ω·s. The plural of henry may be spelled henrys or henries.

Conceptually, it is not obvious that we can equate the electromotive force or voltage *produced* by the inductor, \mathcal{E} in Eq. (3.11), with the voltage *applied* to the inductor by some external source in the circuit, $v(t)$ in Eq. (3.8). But physically, there can only be one unique voltage measured across an inductor at any given same instant. Therefore, \mathcal{E} and v must be the same—even if there is some sleight of hand about the causal relationship between voltage and current.¹⁴ For practical circuit analysis, we needn't worry about whether to interpret the voltage as a cause or an effect of the current. Assuming our model holds, the voltage and current across an inductor at any instant must obey the stated relationship, regardless of cause.

By plugging in some basic sinusoidal functions, we can see how the defining relationship in Eq. (3.8) is consistent with the current lagging the voltage by 90° . Without using any phase angles, we could write $v(t) = V_{\max} \cos(\omega t)$ and $i(t) = I_{\max} \sin(\omega t)$, which makes them 90° apart, and we know from calculus that the derivative of a sine is cosine. Because we are very interested in phase shifts, and they will not always be 90° , we prefer to write both functions in matching form (e.g., both as a cosines) with an explicit phase angle.

For the situation illustrated, we could write $v(t) = V_{\max} \cos(\omega t)$ and $i(t) = I_{\max} \cos(\omega t - 90^\circ)$ or $i(t) = I_{\max} \cos(\omega t - \frac{\pi}{2})$. Note that because current and voltage are in different units, the vertical scales of the two curves are arbitrary.

Because the induced current is related to the change in the magnetic field per unit time, the a.c. frequency is important. The higher the frequency, the more rapidly the magnetic field changes and reverses, and therefore the greater the induced opposed current with its impeding effect. Conversely, the lower the frequency, the easier it is for the current to pass through an inductor.¹⁵

A direct current corresponds to the extreme case of zero frequency. When a steady d.c. voltage is applied to an inductor, it basically behaves like an ordinary piece of wire. After an initial transient period during which the field is established, the magnetic field will remain constant along with the current.¹⁶ An unchanging magnetic field exerts no further influence on an electric current, so the flow of a steady direct current through a coil of wire is unaffected by the inductive property.

As the magnetic field inside an inductor increases and decreases during different parts of the cycle, it stores and releases energy. Specifically, the amount of energy stored in an inductor can be quantified as

$$E = \frac{1}{2}LI^2 \quad (3.12)$$

where E is in joules if L is in henrys and I in amperes. This energy represents the work done in the process of establishing the current through the inductor, and Eq. (3.12) assumes that the inductance was constant throughout that process.¹⁷ This energy is not being dissipated, only

¹⁴ In truth, all the physical devices in a circuit mutually interact, with mutual causality. But their web of microscopic, quantum-mechanical interactions occurs on a much faster time scale and smaller spatial resolution than the circuit behaviors we are concerned with. Even our time-dependent variables like $v(t)$ and $i(t)$ describe phenomena that have, in a sense, already equilibrated, and we use the word “instantaneous” with pragmatic but not mathematical precision. Equations like Ohm’s law, Eqs. (3.8) and (3.14) state requirements for observable quantities that are guaranteed to be met, unless some unusual circumstances render our choice of macroscopic variables inappropriate representations of reality.

¹⁵ The discriminating response of inductors to different frequencies is put to use in electronics. Electronic signals generally contain a multitude of frequencies. When such a signal is applied to an inductor, the lower frequencies are conducted preferentially. For this reason, inductors are also referred to as “low-pass filters.”

¹⁶ In electronics applications, that transient time of “charging up” the inductor can be significant compared to the circuit behavior of interest—in fact, we might deliberately use a large enough inductor to never quite reach a steady state before a switch gets flipped somewhere—and we would have to analyze the circuit in the time domain, using $\frac{di}{dt}$.

¹⁷ Exceptions to this situation include saturated magnetic cores, whose analysis is beyond our scope.

repeatedly exchanged between the inductor's magnetic field and the rest of the circuit. This exchange process becomes very important in the context of power transfer throughout an a.c. circuit, because we will need to account for its precise timing.

Although we used a coil of wire to introduce inductance, the same principles also apply to objects other than coils. The coil makes for the strongest inductance, because the contributions to the magnetic field from different sections of wire add together in an obvious direction. But all conducting objects have *some* inductance, depending on their shape—even straight wires. Section 9.1 discusses inductance of transmission and distribution lines, which turns out to be very significant and impactful for electric grid operations.

3.3.2 Inductive Reactance

The effect of an inductor on an a.c. circuit is expressed by its *reactance*, denoted by X and measured in units of ohms. There are no special varieties of ohms to distinguish reactance from resistance.

To specify inductive (as opposed to capacitive) reactance, we may add the subscript L . The inductive reactance X_L is the product of the angular a.c. frequency and the inductance:

$$X_L = \omega L \quad (3.13)$$

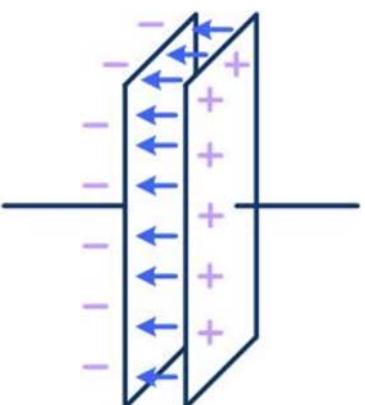
Unlike resistance, and unlike inductance *per se*, reactance is not just determined by the intrinsic characteristics of a device: it also depends on how the circuit is being operated. However, in the context of power systems, where we assume frequency always remains the same to a good approximation, reactance is usually treated as a fixed property of any given object.

3.3.3 Capacitance

Capacitive reactance is, mathematically, a perfect opposite of inductive reactance. Like inductance, capacitance occurs in just about all conducting objects, depending on their shape and sometimes on their interaction with the ground. We focus here on the canonical example, the parallel-plate capacitor. A capacitor may not look like the opposite of a coil, but its effect on the relative timing of current and voltage is exactly backwards.

A basic capacitor consists of two conducting surfaces or plates that face each other and are separated by a small gap (Figure 3.6). These plates can carry an electric charge; specifically, their charges will be opposite. By having an opposite charge on the opposing plate, very nearby but not touching,

Figure 3.6 A basic capacitor, with arrows indicating the electric field.



it is possible to collect a large amount of charge on each plate. We might say that the charge “sees” the opposite charge across the gap and is attracted to it, rather than only being repelled by its like charges on the same plate. In physical terms, there is an electric field across the gap that serves to hold the accumulation of charge on the plates. The gap could simply be air, but is often filled with a better insulating *dielectric* material to prevent sparks from bridging the gap, and allow more charge to be accumulated on the plates. (Even such an insulating material can fail or leak charge; this is known as dielectric breakdown, which will usually ruin a capacitor.) For compactness, the plates are often made out of pliable sheets that are folded or rolled up inside a container.

When presented with a d.c. voltage, a capacitor essentially behaves like a gap in the circuit. Initially, there is a transient period of building up the charge on both plates. But once the charge is built up and cannot go across, the capacitor acts as an open circuit, and no current will flow.

An alternating current, however, can get across the capacitor. Recall from Section 1.1 that although a current represents a flow of charge, individual electrons do not actually travel a significant distance through a conductor; rather, each electron transmits a sort of impulse or “push” to its neighbor. Because this impulse can be transmitted across the gap by means of the electric field, it is not necessary for electrons to physically travel across. This transmission only remains effective as long as the impulse (in other words, the voltage) keeps changing, because once the charge has accumulated on the capacitor plate and a steady electric field is established, there is nothing more to transmit.

Indeed, the current flow across a capacitor is proportional to the rate of change of the electric field, which corresponds to the rate of change of the voltage across the capacitor. As the voltage oscillates, the electric field continually waxes and wanes in alternating directions. The greater the frequency, the more readily the current is transmitted, since the rate of change of the voltage will be greater.¹⁸

This behavior is formalized in the constitutive equation:

$$i(t) = C \frac{dv}{dt} \quad (3.14)$$

The proportionality constant C in Eq. (3.14) is the *capacitance* (not to be confused with the abbreviation C for coulombs), measured in farads (F).¹⁹ We again sidestep the question of causality, and may equally interpret the current being caused by the changing voltage, or *vice versa*.

Like inductance, capacitance depends on an object’s physical shape. Capacitance increases with the area of surfaces where charge is accumulated, and decreases with separation between them (since greater proximity of charges will cause a stronger electric field), as long as there is no contact. For the simple case of two parallel plates of area A separated by distance d with a material in between that is characterized by the dielectric constant k , we can write

$$C = \frac{kA}{d} \quad (3.15)$$

This is stated here mostly for the purpose of jogging memories from introductory physics; we rarely get to apply this tidy equation in the power systems context. Chapter 9 addresses capacitance of transmission lines, which have a more complicated geometry.

Notice the similarity to the constitutive equation for an inductor, but with current and voltage switched. Casually speaking, a capacitor is a thing that resists changes in voltage. This makes sense in that the voltage across a capacitor is related to the amount of charge accumulated on it. In fact,

¹⁸ In electronics, capacitors are therefore used as “high-pass filters.”

¹⁹ To make farads compatible with other SI units, they are quite large; practical capacitors tend to have small fractions of a farad.

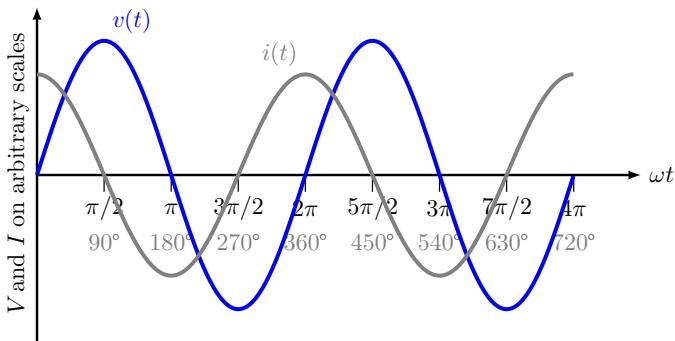


Figure 3.7 Current leading voltage by 90°.

the capacitance C directly relates the amount of charge Q stored on a capacitor's plates to the voltage V across it:

$$Q = CV \quad (3.16)$$

If we remember that current is the flow rate of charge, then we can see that Eq. (3.14) is the result of taking the time derivative of both sides of Eq. (3.16). We can also see that raising or lowering the voltage quickly across a capacitor means delivering or withdrawing a large amount of charge quickly—in other words, a high current.

By comparison to the case of the inductor, we can already anticipate what Eq. (3.14) means for the relative timing of alternating voltage and current through a capacitor: If $v(t) = \sin \omega t$, then $i(t) = \cos \omega t$, and the current is shifted by 90° relative to the voltage—but in the opposite direction as for the inductor! We could say that the voltage lags the current, but in power engineering convention, since we like to take voltage as the reference, we say that the current leads the voltage. In Figure 3.7, if $v(t) = \cos \omega t$, then $i(t) = \cos(\omega t + 90^\circ)$. Notice the graph shows that the alternating current reaching its maximum value at the instant that the voltage changes most rapidly.

Like an inductor, a capacitor stores and releases energy during different parts of the cycle. This energy resides in the electric field between the plates, and is given by

$$E = \frac{1}{2}CV^2 \quad (3.17)$$

Notice this expression is analogous to the inductor energy in Eq. (3.12), with voltage the *dual* of current. The storage and release of energy by a capacitor occurs at time intervals opposite to those of an inductor in the same circuit. This is consistent with the phase shift of current relative to voltage being in opposite directions. A capacitor and an inductor will therefore exchange energy between them in an alternating fashion. Like an ideal inductor, an ideal capacitor only exchanges but never dissipates energy.

3.3.4 Capacitive Reactance

Recall that the inductive reactance X_L comprises information about an intrinsic property of an inductive element (namely, its inductance L) along with the a.c. frequency ω that determines the effect of L on the circuit. Likewise, the capacitive reactance, denoted by X_C , comprises information about both the intrinsic capacitance C of a capacitive element, and ω that determines its effect on

the circuit. However, we may not simply replace the L with a C in Eq. (3.13). Instead, we will write capacitive reactance as

$$X_C = -\frac{1}{\omega C} \quad (3.18)$$

The somewhat unnerving minus sign in Eq. (3.18) expresses the fact that inductive and capacitive reactances have opposite effects on a circuit. The minus sign makes better sense within the complex representation, where it follows from the rules for manipulating imaginary numbers (Section 3.3.5). There are two different conventions: the minus sign can either be subsumed within X_C , as we choose to do in Eq. (3.18), or placed in front of X_C when it appears in the complex impedance \mathbf{Z} as the imaginary jX . Either way, capacitive reactance X_C must be subtracted from, not added to, inductive reactance X_L . In complex notation, we can write

$$\mathbf{Z}_C = jX_C = \frac{1}{j\omega C} = -\frac{j}{\omega C}$$

Irrespective of sign, the crucial take-away from Eq. (3.18) is that the magnitude (absolute value) of the capacitive reactance decreases as the frequency ω increases, and as the capacitance C increases. This is intuitive because an increase in capacitance means that the capacitor plates are becoming more effective at accumulating charge, and thus better at supporting an electric field to transmit a signal across the gap. This makes it *easier* for a current to pass, which corresponds to a *decrease* in reactance. Likewise, at higher frequency, transmitting a current signal becomes easier.

Comparing Eq. (3.18) to the inductive case, it is helpful to refer to *dual* concepts (initially introduced in Section 2.2.4), now extended to electricity and magnetism, and the complex admittance introduced in Section 3.3.7. The reader may wish to revisit this explanation later; it is kept adjacent to the capacitance section for reference.

Because electricity and magnetism are two complementary ways to observe what is fundamentally the same physical phenomenon, there are always two complementary or dual mathematical formulations of any true statement about electromagnetism: one in terms of electricity, and one in terms of magnetism. Capacitance is the dual of inductance, meaning that a capacitor behaves electrically analogous to the way an inductor behaves magnetically. When rewriting a statement in terms of its dual concept, we must be careful to switch *all* the dual terms.

Importantly, reactance has a dual term called *susceptance*—just as the dual of resistance is conductance. While we think of an inductor as adding a reactance in series, the most useful way to think of a capacitor is that it adds a susceptance in parallel. Rather than making it more difficult for current to pass, it provides an additional path to make it easier.

In this light, we could write the corresponding relationship in terms of the *susceptance*, which is denoted by the symbol B (not to be confused with the magnetic field) and measured in units of siemens (S) like conductance. For the record,

$$B_C = \omega C$$

Instead of referring to susceptance B , capacitors are often described by the more general term *admittance*, symbol Y , which includes both susceptance and conductance in parallel (see Section 3.3.7). Capacitors physically resemble their idealized versions very closely, meaning they include only susceptance and no conductance, since even a small imperfection in a capacitor would render it useless.²⁰ When there is zero conductance, admittance and susceptance

²⁰ This situation is different from an inductor, which ideally consists only of reactance, but in practice may include significant amounts of resistance in series.

become effectively interchangeable. In terms of the complex admittance \mathbf{Y} and its dual or inverse *impedance* \mathbf{Z} , we can write:

$$\mathbf{Y}_C = jB_C = j\omega C = \frac{1}{Z_C} = \frac{1}{jX_C} = -j\frac{1}{X_C}$$

The introduction of complex numbers is now overdue.

3.3.5 Complex Numbers

This section can be skipped by readers familiar with complex arithmetic. It is included here as a prerequisite for defining the complex impedance and introducing phasor notation, recognizing that academics and professionals in most fields outside of mathematics, physics or electrical engineering might never have had occasion to work with complex numbers.

Complex numbers are a concise way to mathematically represent two aspects of a physical system at the same time. This will be necessary for describing electrical impedance as a combination of resistance and reactance in the following section. A complex number contains a real part, which is an ordinary number and directly corresponds to a measurable physical quantity, and an imaginary part, which is a sort of intangible quantity which, when projected onto physical reality, is associated with oscillatory behavior.

An imaginary number is a multiple of the imaginary unit quantity $\sqrt{-1}$. This quantity is denoted by i for imaginary in mathematics, and j in electrical engineering, so as to avoid confusion with the label for current. The definition $j = \sqrt{-1}$ is another way of saying that $j^2 = -1$. It also implies that $1/j = -j$.

This entity j embodies the notion of oscillation, or time-varying behavior, in its very nature. Consider the equation $x^2 = -1$. There is no real number that can work in this equation if substituted for x . Pick any positive number for x , and x^2 is positive. Pick any negative number for x , and when you square it, the result is also positive. Pick zero, and x^2 is zero. So we devise an abstract object that we call j —not a real number as we know it, but a thing which, it turns out, can also be manipulated just like a regular number. We define j as that thing that makes the equation $x^2 = -1$ true.

You can think of j as the number that cannot decide whether it wants to be positive or negative. In fact, the equation $x^2 = -1$ can be translated into the logical statement, “This statement is false.” The statement cannot decide whether it is true or false; it flip-flops back and forth. In essence, the imaginary j is just “flippety.”²¹

The number j makes the most sense when we represent it graphically. Consider the number line with positive and negative real numbers, the positive numbers extending to the right and the negative numbers to the left of zero. Now think of multiplying a positive number by -1 . What does this operation do? It takes the number from the right-hand side of the number line and drops it over to the left. For example, $3 \cdot -1 = -3$. In other words, we can think of the multiplication by -1 as a rotation by 180° about the origin (the zero point). Now we can take the result (say, -3) and multiply it again by -1 . We get $-3 \cdot -1 = 3$. In other words, the number was again rotated by 180° . Successive applications of the “multiplication by -1 ” operation result in successive rotations.

So far, we have restricted our imagination to numbers that lie on the real number line. Now suppose we “think outside the line” and ask the following question: What if there were an operation

²¹ I owe this term to Heinz von Foerster, as rendered in the transcript of the 1973 AUM conference at Esalen Institute (*Laws of Form: Spencer-Brown at Esalen 1973, Cybernetics & Human Knowing*, Vol. 26 No. 2-3, 2019) referring to G. Spencer Brown, *Laws of Form* (New York: Crown Publishing, 1972).

that rotates a number not by 180° , but by only 90° ? Rotating by 90° does not immediately seem meaningful, because it takes us off the number line into uncharted territory. But we do know this: if performed twice in succession, a rotation by 90° corresponds to a complete 180° rotation. In other words, a rotation by 90° , performed twice, gives the same result as multiplying by -1 .

But this leads us directly to the equation $x^2 = -1$, which asks for the number that, when multiplied by itself, becomes negative. If we define “multiplying by j ” as “rotating by 90° ” and “multiplying by j^2 ” is the same as “multiplying by j and then multiplying by j again,” then to multiply by j^2 is to rotate by 180° , which is exactly the same thing as multiplying by -1 . In this sense we can say that j^2 and -1 are the same.

Based on this rotation metaphor, j is conceptualized as the number that is measured in the novel “imaginary” direction at right angles to the real number line. We can now extend this concept of being “off the real number line” to an entire new, imaginary number line orthogonal to the real one, intersecting at the origin or zero point. Along this new axis we can measure multiples of j upward and $-j$ downward. By convention, positive angles are measured in the counterclockwise direction, so $+90^\circ$ takes us to $+j$ pointing upward.

With these two directions, we have in effect converted the one-dimensional number line into a two-dimensional plane of numbers, called the *complex plane*. This plane is defined by a real axis labeled *Re* (the old number line) and an imaginary axis labeled *Im* (the new imaginary number line). We can now conceive of numbers that lie anywhere on this complex plane and represent a combination of real and imaginary numbers. These are the complex numbers. The complex number $C = a + jb$ (not to be confused with capacitance or coulomb) refers to the point a units to the right of the origin and b units above. We say that a is the real part and b is the imaginary part of C . In Figure 3.8, $a = 3$ and $b = 4$.

Rather than specifying the real and imaginary components explicitly, another way of describing the same complex number is in reference to an arrow (vector) drawn from the origin to the point corresponding to the number. The length of the arrow is the magnitude of the complex number. This magnitude is a real number and may be denoted by vertical lines, as in $|C|$. To capture all the information about C , the angle between the arrow and the real axis is also specified, which we denote by $\angle\theta$ (Greek lowercase theta). The representation in terms of magnitude and angle is called the *polar form*, as opposed to rectangular (real plus imaginary). These alternative representations

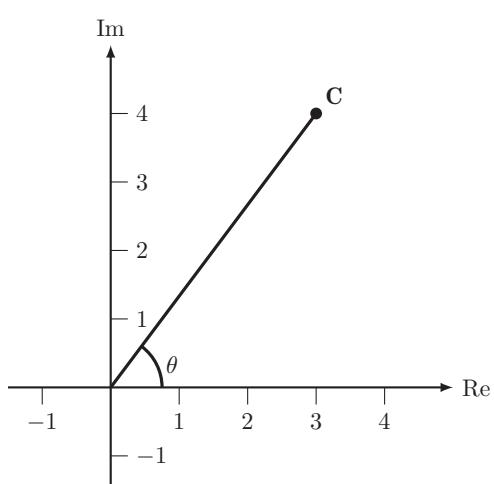


Figure 3.8 The number $C = 3 + j4$ in the complex plane. In polar notation, $C = 5\angle53.1^\circ$.

can be converted into one another by using any two of the following relationships:

$$|C|^2 = a^2 + b^2 \quad \sin \theta = \frac{b}{|C|} \quad \cos \theta = \frac{a}{|C|} \quad \tan \theta = \frac{b}{a}$$

Adding or subtracting complex numbers is easily done in the rectangular format: we simply combine the real parts and the imaginary parts, respectively. To multiply two complex numbers, both components of one are multiplied by both components of the other, and the results are added (like in a common binomial expression). However, multiplying, and especially dividing, complex numbers is much easier in the polar format: the magnitudes are multiplied together (or divided), and the angles are added (or subtracted). Adding the angles is consistent with our earlier discussion of multiplication as successive rotations. For division, the divisor angle is subtracted. In the complex plane, subtracting an angle appears as a backwards or clockwise rotation, essentially “undoing” a multiplication.

A commonly used term with complex numbers is the *complex conjugate*, denoted by an asterisk. This means simply taking the negative of the imaginary part, or the negative of the angle. For example, if $C = 3 + j4 = 5\angle 53.1^\circ$, then its complex conjugate is $C^* = 3 - j4 = 5\angle -53.1^\circ$. Graphically, the complex conjugate is a reflection about the real axis. It is distinct from the *inverse* $1/C = 1/5\angle -53.1^\circ$, and from $-C = -5\angle 53.1^\circ = 5\angle 233.1^\circ$.

The imaginary j and all the complex numbers that spring from it do not have the same utilitarian properties that real numbers do. You cannot eat j eggs for breakfast, and you cannot be the j th person in line at the post office. Nevertheless, complex numbers as operational devices do obey rules of manipulation that qualify them as “numbers” in the mathematical sense. For instance, complex numbers can be used in exponentiation. Their special properties in these manipulations make them useful tools for representing certain real phenomena—especially ones that involve “flippety,” like alternating current.

3.3.6 Complex Impedance

The combination of reactance and resistance that describes the overall behavior of a device in a circuit is called the *impedance*, denoted by Z . However, Z is not a straightforward arithmetic sum of R and X . Mathematically speaking, Z is the vector sum of R and X in the complex plane. All three quantities have the same units of ohms.

A boldface \mathbf{Z} may be used to remind us that we are dealing with a vector or complex quantity, or absolute value signs may be used to indicate the magnitude. These formatting conventions are not always applied conscientiously (even within this book), and power engineers are often expected to infer from context whether a variable is real-valued or complex.

Figure 3.9 illustrates the impedance \mathbf{Z} as a complex number whose real part is the resistance and whose imaginary part is the reactance:

$$\mathbf{Z} = R + jX \tag{3.19}$$

Any device found in an electric power system has an impedance. For different devices under different circumstances, either the resistive or reactive component may be negligible, but it is never wrong to use the more inclusive Z instead of just R or X alone.

Impedances can be combined according to the same rules for series and parallel combination that we showed in Section 2.2 for pure resistances, but the arithmetic with complex numbers is more challenging. When combining circuit elements in series, we add their impedances. When combining them in parallel, we add their *admittances*, or inverse of impedance.

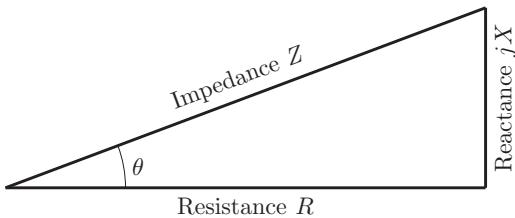


Figure 3.9 The impedance Z represented in the complex plane, with resistance R in the real direction and reactance jX in the imaginary direction. Here the reactance is inductive and points upward; a capacitive reactance would point downward in the direction of $-j$. The length of the hypotenuse is $|Z|$.

Keep in mind that systems with alternating current are two-dimensional, in the sense that they are always described by two variables. When studying the effect of an impedance on a circuit with a given voltage source, we are interested in two things: magnitude and timing. The magnitude of the impedance affects the magnitude of the current relative to the voltage, and its angle affects the timing of the current relative to the voltage. This is better seen with the complex impedance in polar form:

$$Z = |Z|\angle\theta \quad (3.20)$$

where

$$|Z| = \sqrt{R^2 + X^2} \quad \text{and} \quad \tan\theta = \frac{X}{R} \quad (3.21)$$

The angle $\angle\theta$ of the impedance has an important physical significance: it corresponds to the phase shift between current and voltage produced by this circuit element. Crucially, the same angle that represents time in the context of the sine wave can also be viewed spatially in the complex plane, where we map the impedance as a vector. Figure 3.9 shows the impedance with a horizontal real component representing resistance, and a vertical imaginary component. The imaginary component may point up or down, depending on whether the reactance is inductive or capacitive. This is where the minus sign from Eq. (3.18) may begin to make sense.

By convention, we take a positive angle θ to denote a current *lagging* behind voltage, as shown in Figure 3.5. The lagging current is caused by an inductive reactance X_L , which gives the impedance a positive imaginary part. Conversely, a negative angle θ indicates a current leading voltage (as in Figure 3.7), due to a capacitive reactance X_C that we have defined to be negative, and therefore gives the impedance a negative imaginary part.

When circuit elements are placed in series, their impedances add. A single physical object can have both a resistive and a reactive property, and the combination of the real and imaginary parts of its impedance is an addition process. For example, a non-ideal inductor made from a coil of copper wire, which has some resistance as well as some inductive reactance, is represented as an ideal inductor plus a resistor in series. In our standard circuit modeling context, a series of sequential ideal elements is treated equivalent to a commingled situation.

When we combine inductive and capacitive elements in a circuit, their reactances tend to counteract or compensate for each other. In fact, a circuit with a perfectly matched inductance and capacitance could appear from the outside as having no reactance at all. However, the match only holds true at one specific frequency ω , where $|X_L| = |X_C|$.²² This corresponds to the physical situation where the inductor and capacitor exchange energy with each other, independent of the rest of the circuit.

Reactances may exchange energy whether they are in series or in parallel. In the series case, it is straightforward to compare reactances. The analysis of parallel elements is better done using admittances, as in the following section.

²² Readers familiar with electronics will recognize this as a resonance condition.

Example

An electrical device contains a resistance, an ideal inductance and an ideal capacitance, all connected in series. Their values are $R = 1 \Omega$, $L = 0.01 \text{ H}$ and $C = 0.001 \text{ F}$, respectively. At 60 Hz, what is the total impedance?

The impedance is the complex sum of the resistance and both reactances: $\mathbf{Z} = R + jX_L + jX_C$.

The magnitude of the inductive reactance is

$$|X_L| = \omega L = 377 \text{ rad/s} \cdot 0.01 \text{ H} = 3.77 \Omega$$

To indicate that this quantity is reactive, it is helpful to make the j explicit and write $jX_L = j3.77 \Omega$.

The magnitude of the capacitive reactance is

$$|X_C| = \frac{1}{\omega C} = (377 \text{ rad/s} \cdot 0.001 \text{ F})^{-1} = 2.65 \Omega$$

To account for the negative imaginary direction, it is least ambiguous to write $jX_C = -j2.65 \Omega$.²³

The impedance is the complex sum of the resistance and both reactances: $\mathbf{Z} = 1 + j3.77 - j2.65 \Omega = 1 + j1.12 \Omega$.

In polar form, this corresponds to $\mathbf{Z} = 1.5 \angle 48.28^\circ \Omega$.

The positive angle of the impedance indicates that the inductive property is dominant. From outside the device, we don't observe its internal capacitance at all. Rather, it behaves like a resistor and a 1.12Ω inductor.

Casually speaking, this impedance might be simply referred to by its magnitude, 1.5Ω , without mention of the angle. This would tell us about the magnitude but not the timing of current relative to an applied voltage.

3.3.7 Complex Admittance

The inverse of the complex impedance is the *admittance*, denoted by \mathbf{Y} and measured in units of siemens (S). The complex \mathbf{Y} has real and imaginary parts called the *conductance* G and the *susceptance* B (no relation to the magnetic field of the same letter):

$$\mathbf{Y} = G + jB \tag{3.22}$$

For the special case of a pure resistor without inductance in Section 1.2.2, we already encountered the conductance as the inverse of resistance. More generally, the complex admittance is the inverse of the complex impedance. The inverse of a complex number is most easily taken in polar form:

$$\mathbf{Y} = \frac{1}{\mathbf{Z}} = \frac{1}{|Z| \angle \theta} = \frac{1}{|Z|} \angle -\theta \tag{3.23}$$

Note that the admittance magnitude is the inverse of the impedance magnitude, and that the admittance angle in the complex plane is the negative of the impedance angle.

In rectangular form, this inverse is much trickier. It is especially important to realize that for the complex case, $G \neq 1/R$, and $B \neq 1/X$. Some clever algebra²⁴ yields the following expression:

$$\mathbf{Y} = \frac{1}{\mathbf{Z}} = \frac{1}{R + jX} = \frac{(R - jX)}{(R + jX)(R - jX)} = \frac{R}{R^2 + X^2} - j \frac{X}{R^2 + X^2} \tag{3.24}$$

²³ In the alternative style convention, we could keep X_C positive, but write $\mathbf{Z} = R + jX_L - jX_C$. Luckily, this does not tend to cause confusion in practice, since it is understood that the capacitive reactance ultimately needs to be opposite the inductive reactance, regardless of where the minus sign or the j are kept. It is also not uncommon, if mathematically inaccurate, to see the j subsumed within the reactance, as in $X_C = -j2.65 \Omega$.

²⁴ A standard trick for dealing with complex fractions is to multiply both numerator and denominator by the complex conjugate of the denominator. This conveniently eliminates the denominator's imaginary part, so that we may separate the real and imaginary parts of the whole expression.

Therefore, we can write for the real and imaginary parts of Y :

$$G = \frac{R}{R^2 + X^2} \quad \text{and} \quad B = -\frac{X}{R^2 + X^2} \quad (3.25)$$

Now we can see that for the purely resistive case where $X = 0$, G does in fact equal $1/R$ and B goes away, while for the purely reactive case where $R = 0$, G goes away and $B = -1/X$. In other words, a device with a purely real impedance will also have a purely real admittance, and a device with a purely imaginary impedance will have a purely imaginary admittance. It is crucial to recognize that these are special cases.

Generally, the relationship between the magnitudes of G and B (effectively, the angle of Y) is directly proportional to the relationship between R and X (the angle of X , where $\angle X = -\angle Y$). Therefore, a device whose reactance dominates over its resistance also has a susceptance that dominates over its conductance. This is an important qualitative observation, because in practice we often assume that either resistance or reactance dominates for a particular component (so as to simplify our model for its behavior).

Note that switching the sign of the imaginary component is the same as switching the sign of the angle in the complex plane. To summarize: inductive elements have positive X (or $\angle Z$) and negative B (or $\angle Y$), while capacitive elements have negative X (or $\angle Z$) and positive B (or $\angle Y$).

It is worth remembering what the negative sign in the complex plane represents physically: a shift in time. The translation between angle and time comes from having defined the duration of a complete cycle as 360° . Reversing the sign of the angle implies a change between lagging and leading current. The angle of admittance is the same as the angle of current relative to voltage, while the angle of impedance will be of opposite sign. We will confirm this by writing Ohm's law in complex form in Section 3.4 (Eq. (3.44)).

When analyzing circuits, it helps to move nimbly back and forth between impedance and admittance representations. Whenever we consider components in series, we want to add their impedances; in parallel, we want to add their admittances. This is best shown with a simple example.

Example

In Figure 3.10, the series reactance from the inductor and capacitor is partially compensated. The combined series impedance (taking the minus sign to be subsumed within X_C) is:

$$Z_s = R + jX_L + jX_C = 1 + j2 - j0.5 = 1 + j1.5 \quad (\Omega)$$

If instead we combine the same three components in parallel as shown in Figure 3.11, we first convert impedances to admittances (remembering that $1/j = -j$) and then add the parallel admittances:

$$Y_p = G + jB_L + jB_C = \frac{1}{1} + \frac{1}{j2} + \frac{1}{-j0.5} = 1 - j0.5 + j2 = 1 + j1.5 \quad (\text{S})$$

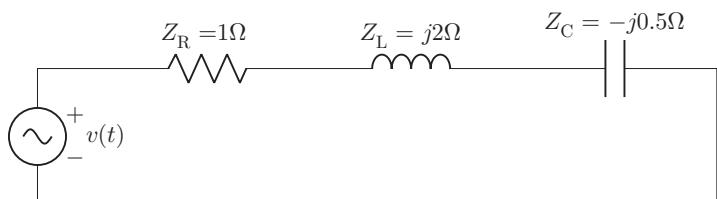


Figure 3.10 Simple series circuit to illustrate the addition of impedances.

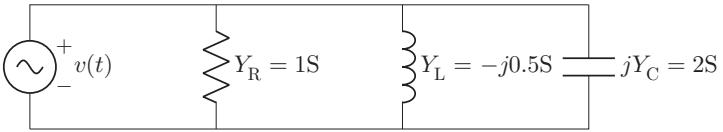


Figure 3.11 Parallel circuit to illustrate the addition of admittances, using the same elements as in Figure 3.10.

If we wish to express the combined behavior of the three elements in terms of an impedance, we may invert the admittance as the last step. This is easiest to do in polar form:

$$Y_p = \sqrt{1 + 1.5^2} \angle(\tan^{-1}1.5) = 1.8\angle56.3^\circ (\text{S})$$

$$Z_p = \frac{1}{Y_p} = 0.555\angle-56.3^\circ (\Omega)$$

Let's compare the behavior of these two circuits. The series impedance, expressed in polar form, is $Z_s = 1.8\angle56.3^\circ$. When we take the same three physical elements in parallel, their combined impedance will be less than in series ($|Z_p| < |Z_s|$). This is intuitive because more paths make it easier for current to flow.

Note how these same elements combine to present an overall inductive or capacitive load, depending on how they are connected. In the series combination, the effect of the inductor dominates over the capacitor, as it presents a four times greater impedance; consequently, the combination draws an overall lagging current, as indicated by the positive angle of Z_s . In the parallel combination, however, the capacitor dominates, making the angle of Z_p negative. Intuitively, this is because its greater admittance allows more current to flow through its separate circuit branch. With more current through the capacitor, its effect on the circuit is more pronounced, and the total current supplied by the source is leading. We will quantify these observations after introducing complex power.

One last observation before we leave this example: Parallel branch current and series voltage drop are dual concepts. Sharing the same voltage source, the branch currents will be proportional to the admittances. Thus, the resistor will draw $2/7$, the inductor $1/7$, and the capacitor $4/7$ of the combined current. In the series case, the voltage drop across the resistor will be $2/7$, the inductor $4/7$, and the capacitor $1/7$ of the source voltage, proportional to the impedance of each element. These statements can be formalized with Ohm's law using complex voltages and currents.

3.4 Complex Power

We wish to quantify the power conveyed in an a.c. circuit that includes reactance as well as resistance. I choose the word “conveyed” because it is agnostic about direction, and whether power is dissipated or merely exchanged within the circuit. The time shift between current and voltage introduced by reactance will crucially affect the direction of power transfer, but the reader shall be forewarned that the common nomenclature is imprecise and can be confusing.

As noted earlier (Eq. (3.4)), it is always correct to write $p(t) = i(t)v(t)$ for instantaneous power. However, we now want to find appropriate expressions that average over one or more a.c. cycles. The elegant and concise way to adapt the d.c. equation $P = IV$ to the a.c. environment is to convert

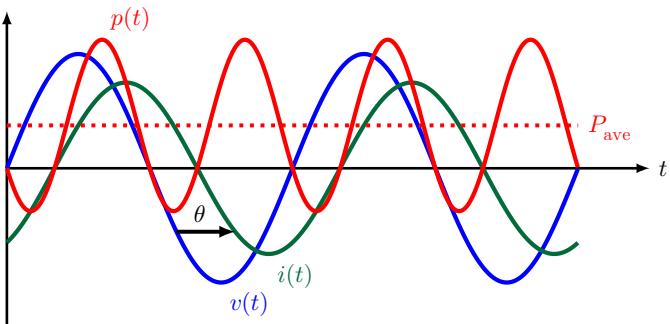


Figure 3.12 Power as the product of voltage and current, with current lagging behind voltage by a phase angle difference θ .

all three terms into complex or *phasor* quantities (Section 3.5) and write $\mathbf{S} = \mathbf{I}^* \mathbf{V}$. But first, let us focus explicitly on the product of voltage and current waveforms in the time domain.

When inductance or capacitance causes a shift in the relative timing of voltage and current, one quantity is sometimes negative when the other is positive. As a result, the product of voltage and current—that is, the instantaneous power—is sometimes negative. This is shown in Figure 3.12, where $p(t)$ dips below the horizontal axis. Physically, the relative direction of voltage and current tell us whether the electrons are doing work or having work done on them: if a charge is moving toward a lower potential energy, it is doing work on its surroundings; if it is moving toward a higher potential, it is being worked on or absorbing energy. We can interpret the negative instantaneous power as saying that power flows “backwards,” or out of the load and back into the source or generator, for a brief interval during each half cycle. The energy that is being transferred back and forth is stored alternately in electric or magnetic fields within these devices.

Since instantaneous power is sometimes negative, the average power flowing into the load is clearly less than it was in the resistive case. But just how much less? We know what to expect in the purely reactive scenario: if the phase shift is a full 90° in either direction, positive and negative power will cancel exactly, and the average power will be zero. This is the case of the ideal inductor or capacitor, which dissipate no energy. Conversely, if the phase shift is very small, we have the resistive case. For anything in between, the reduction in average power is not visually obvious from a graph of $p(t)$, but is in fact quite straightforward to determine mathematically.

3.4.1 Real Power and Power Factor

For time-shifted sinusoidal waveforms as described in this chapter, the average power is proportional to the cosine of the phase shift between the voltage and current waveforms. That cosine is called the *displacement power factor*, because it is associated with current being displaced in time from voltage. Using θ to denote the phase angle difference between voltage and current, we can write

$$P_{\text{ave}} = VI \cos \theta \quad (3.26)$$

Let us derive this result by revisiting the expression for power from the resistive case earlier, using the distinct phase shift ϕ_V and ϕ_I for voltage and current, respectively. The following expression

holds in general, for any kind of impedance:

$$p(t) = v(t)i(t) = V_{\max} \cos(\omega t + \phi_V) I_{\max} \cos(\omega t + \phi_I) \quad (3.27)$$

Here we face the task of multiplying cosines with different arguments, prompting us to dig deeper into the stores of helpful trig identities. We find that $\cos \alpha \cos \beta = \frac{1}{2}[\cos(\alpha - \beta) + \cos(\alpha + \beta)]$ lets us write:

$$p(t) = \frac{1}{2}V_{\max}I_{\max} \cos(\phi_V - \phi_I) + \frac{1}{2}V_{\max}I_{\max} \cos(2\omega t + \phi_V + \phi_I) \quad (3.28)$$

Since voltage and current have the same frequency, ωt goes away in the difference between arguments, and the first term in Eq. (3.28) no longer contains any time variation. The second term, with twice the original frequency (plus a phase shift that is inconsequential), varies between $\pm \frac{1}{2}V_{\max}I_{\max}$. The total instantaneous power transfer will therefore dip below zero somewhere during the cycle if $\phi_V \neq \phi_I$.

But we are primarily interested here in the average power over one or more cycles. For this, we can simply discard the oscillating second term, which by itself averages to zero. We are left with only the constant term, where we label the angle difference as θ and adopt root-mean-square magnitudes:

$$P_{\text{ave}} = \frac{1}{2}V_{\text{rms}}I_{\text{rms}} \cos(\phi_V - \phi_I) = V_{\text{rms}}I_{\text{rms}} \cos(\phi_V - \phi_I) = VI \cos \theta \quad (3.29)$$

The displacement power factor $p.f.\cdot_{\text{disp}}$ is defined as

$$\cos \theta \equiv p.f.\cdot_{\text{disp}} \quad (3.30)$$

Notice that the angle $\theta = (\phi_V - \phi_I)$ is precisely the angle of the complex impedance Z in polar coordinates. This is an important insight: the load impedance is what determines the power factor for a circuit. For the purely resistive case, $\phi_V = \phi_I$ and $\theta = 0$, so that $\cos \theta = 1$ and $P_{\text{ave}} = VI$, consistent with Eq. (1.4).

Typically, where ϕ_V and ϕ_I are in the same quadrant, the power factor will be some number between 0 and 1: zero for an ideal inductor or capacitor, and 1 (unity) for a pure resistor. Notice that the sign of the cosine does not flip depending on whether current lags or leads voltage. Therefore, we need to use the terms “lagging” and “leading” for specifying the displacement power factor. Only when the shift between current and voltage is greater than 90° does average power become negative.

The average power P_{ave} is also called *real power* or *active power*, since it corresponds to the net energy transferred into the load over time. We will drop the subscript and simply write P for real power, well aware that P alone does not tell the whole complex story.

Colloquially, displacement power factor is often taken to be synonymous with “power factor,” although this can be inaccurate and misleading. Power factor, more generally defined as the ratio of real to apparent power (Section 3.4.3), also has a component associated with *harmonic distortion*, or non-sinusoidal waveforms. In the interest of wrangling one new concept at a time, we postpone distortion until Section 5.3 and consider the shorthand “power factor” in the present chapter to mean exclusively “displacement power factor.” This amounts to assuming all voltages and currents to be ideal sinusoidal functions.

3.4.2 Reactive Power

Next, we wish to account for the oscillating portion of instantaneous power. However, we are not simply interested in the fact that power dissipated in a load waxes and wanes over the course of a cycle. Rather, we care to describe the portion of the power that dips below zero, meaning that

it is actually exchanged with circuit devices elsewhere. This corresponds to the power associated with purely inductive or capacitive elements. We will call this quantity *reactive power*, symbol Q , measured in units of volt-ampere reactive (VAR, VAr, or var).²⁵

Reactive power is considered a problematic concept by many engineers. The quantity Q has no single “correct” definition or physical identity like instantaneous power $p(t) = v(t)i(t)$. Rather, Q is based on a convention²⁶ for combining and averaging measurable quantities under certain simplifying assumptions (which made it tractable with 19th and 20th century techniques). Reactive power carries useful and actionable information if and when conditions in a real power system approximate these assumptions well enough.

With these caveats, reactive power is given by

$$Q = VI \sin \theta \quad (3.31)$$

mirroring Eq. (3.26) for P and using the same variables.

Equation (3.31) is obtained by extracting Q from the expression for instantaneous power in Eqs. (3.27) and (3.28), using another trigonometric identity: $\cos(\alpha - \beta) = \cos \beta \cos \alpha + \sin \beta \sin \alpha$. For simplicity, let’s choose the voltage phase angle to be zero (no generality is lost, since what matters is the angle difference between voltage and current). We rewrite Eq. (3.27) for instantaneous power in simplified notation with rms values, θ , and letting $\phi_V = 0$:

$$p(t) = VI \cos \theta + VI \cos(2\omega t - \theta) \quad (3.32)$$

Now we apply the above trig identity to the term with the difference in the argument:

$$p(t) = VI \cos \theta + VI \cos \theta \cos 2\omega t + VI \sin \theta \sin 2\omega t \quad (3.33)$$

This rearrangement has cleverly decomposed the instantaneous power into three terms: the constant average power, plus two distinct oscillating terms that play very different roles.

Notice that regardless of the power factor (assuming only that $\theta \leq 90^\circ$), the sum of the first two terms in the equation never becomes negative; at the minimum when $\cos 2\omega t = -1$, the sum just hits zero. In fact, the first two terms combined reflect the power dissipated by the resistance alone. In the case where the impedance is purely resistive, $\cos \theta = 1$ and $\sin \theta = 0$, making the third term disappear.

The third term in Eq. (3.33) applies strictly to the reactive portion of the circuit. The $\sin 2\omega t$ oscillation always averages to zero, meaning that no net energy has been gained or lost. We can also say that the reactive load draws a current of magnitude $I_X = I \sin \theta$, which is 90° out of phase with I_R (from the shift between $\sin 2\omega t$ and $\cos 2\omega t$). If the impedance is either purely inductive or capacitive, $\sin \theta = 1$ and $\cos \theta = 0$, leaving only the third term.

Notice that we have effectively separated the waxing and waning of positive instantaneous power from that portion of power that becomes negative and is exchanged with the rest of the circuit. The amplitude of that latter oscillating power is $VI \sin \theta$, which represents the *reactive power* Q .

The key observation is that reactive power is determined by the load impedance with its angle $\theta = (\phi_V - \phi_I)$. But unlike the cosine, which stays positive regardless whether current lags or leads, the sine changes sign. Thus, depending on whether a circuit element is capacitive or inductive, it

²⁵ Since “reactive” is only a modifier of the proper abbreviations V and A, the notation VAr is most logical, but arguably least pleasing to the eye. Many academic texts prefer var, while VAR is standard in industry. Take your pick, but don’t ever write Var.

²⁶ Some might say that reactive power is a socially constructed quantity.

will exchange Q with a particular timing that is characterized by convention as a positive or negative contribution to the circuit.

A circuit element that behaves inductively, with positive Q and θ , is said by convention to “consume” reactive power, while a capacitive element is said to “produce” or “generate” Q . From a theoretical or conceptual standpoint, this power engineering vernacular is gravely misleading. It came about because a majority of traditional electric loads happen to be inductive in nature, so that a lagging current is *by circumstance* associated with real power P being consumed.²⁷

It cannot be overemphasized that the word choice of “producing” and “consuming” Q is based on historical convention, not physics. There is no physical dissipation of reactive power; it does not leave the electric circuit, and in that sense is never “consumed.” There is only an exchange between inductive and capacitive elements: at the instant that the inductor magnetic field absorbs energy, the capacitor electric field (or the magnetic field in a generator with opposite timing that “produces” VARs) in the same circuit releases energy. Conversely, at the instant that the inductor magnetic field releases energy, another field somewhere absorbs it. Although on average neither inductive nor capacitive element gains or loses energy, their effects are complementary—like children on a seesaw bouncing up and down.

One benefit of the “producing-consuming” nomenclature for reactive power is that it evokes the idea of energy conservation, which is useful in operational practice. For an a.c. circuit to operate in a steady state, it is necessary that reactive power “production” and “consumption,” as time-averaged quantities, are matched. This follows from the physical requirement that instantaneous power into and out of the circuit be equal at all times during each cycle. Thus, if a load “demands” a certain amount of Q , either there is some source in the circuit that “provides” exactly that amount, or it will become impossible to maintain steady-state voltages and currents. Metaphorically, if one end of the seesaw goes up, the other must go down—or the seesaw breaks.

Analogous to the expressions $P = I^2R$ and $P = V^2/R$ for real power, reactive power can also be written as $Q = I^2X$ or $Q = V^2/X$, depending on whether we know V across or I through the inductive device in the given situation. These scalar relationships can be tricky to apply correctly for complex loads and are most useful for very simple cases.

3.4.3 Power in the Complex Plane

The combination of real and reactive power, with their respective factors of $\cos \theta$ and $\sin \theta$, lends itself perfectly to graphing in the complex plane, analogous to the complex impedance. Their vector sum is the *complex power*, symbol \mathbf{S} :

$$\mathbf{S} = P + jQ \quad (3.34)$$

The magnitude S (more carefully written $|S|$) of complex power is called *apparent power*, measured in volt-amperes (VA), and is represented by the hypotenuse in Figure 3.13.

The projection of complex power onto the real axis has length P and corresponds to the real power; the projection of apparent power onto the imaginary axis has length Q and corresponds to reactive power. This “power triangle” is perfectly similar to the impedance triangle in Figure 3.9, with the same angle θ that is determined by the impedance, where P is proportional to R and Q is proportional to X . It is important to remember that all three quantities in Figure 3.13 are scalars,

²⁷ Inductive behavior also tends to dominate throughout transmission and distribution systems. To compensate for both customer loads and infrastructure, generators are recruited by acting in the opposite manner, like a capacitance. Therefore, inductive behavior in electric power systems is *circumstantially* associated with things that consume real power, while capacitive behavior is associated with things that produce real power.

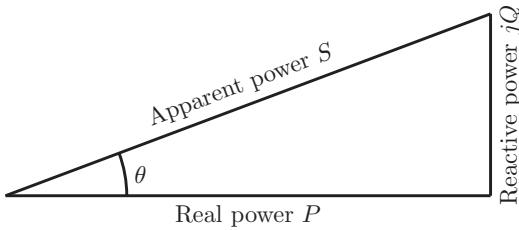


Figure 3.13 Complex power \mathbf{S} with real power P in the real and reactive power Q in the imaginary direction. The positive angle θ in the figure indicates an inductive load, with current lagging voltage.

not vectors. They are represented in a triangle merely for convenience—essentially, as a mnemonic for doing arithmetic—since their magnitudes happen to be related by trigonometric functions.

Again, these relationships only apply in the case of sinusoidal voltages and currents. More generally, the power factor can still be defined as the ratio of average to apparent power,

$$p.f. = \frac{P_{\text{ave}}}{S}$$

In general, that ratio will not be given by a simple cosine, as other phenomena (such as distorted waveforms, Section 5.3.1) may conspire to increase the magnitude of apparent power.

The sign convention for complex power \mathbf{S} is the same as for complex impedance \mathbf{Z} , where a positive angle θ indicates a lagging (inductive) load and a negative θ indicates a leading (capacitive) load. As we will see in Section 3.4.5, this representation allows us to analyze circuits graphically in the complex plane, by combining positive and negative contributions of Q from various circuit elements in the vertical (imaginary) direction.

The apparent power contains information about the total current flowing, irrespective of the current's timing relative to voltage (and thus irrespective of how much real energy is delivered to the load). This total current is what matters in the context of thermal ratings for equipment, where the internal heat produced is I^2R regardless of any phase shift. For this reason, generators and transformers are usually rated in units of apparent power, such as kilo- or megavolt-ampere (kVA or MVA).

Example

Consider a vacuum cleaner that draws 750 W of real power, at a voltage of 120 V a.c. and a power factor $p.f. = 0.75$ lagging. How much current does it draw?

From context, the $p.f.$ is understood to mean the displacement power factor $\cos \theta$, and real power corresponds to $P_{\text{ave}} = V_{\text{rms}} I_{\text{rms}} \cos \theta$. That is, the real power is given by the apparent power times the power factor. The apparent power will inform us about the total current.

The apparent power in this example equals $S = 750 \text{ W} \div 0.75 = 1000 \text{ VA} = 1 \text{ kVA}$. The rms current is the apparent power divided by the rms voltage: $I = S/V = 1000 \text{ VA} \div 120 \text{ V} = 8.33 \text{ A}$.

If we didn't know about the power factor, we might have expected the current to be just 6.25 A. But that would only account for the portion of current that is time-aligned with voltage and contributes to real power. The total 8.33 A current can be understood as the vector sum of 6.25 A, plus a component that lags by 90° (using the result below, we could determine the lagging component is 5.5 A). Because these two current contributions are offset in time, they don't add arithmetically.

How much reactive power does the vacuum cleaner draw?

With $p.f. = 0.75$ lagging, the phase shift is $\theta = \cos^{-1} 0.75 = 41.48^\circ$, and $\sin \theta = 0.661$. Thus, the reactive power is $1000 \cdot 0.661 = 661 \text{ VAR}$. Because the power factor is lagging, the vacuum cleaner is said to “consume” reactive power.

What is the impedance of the vacuum cleaner?

Here we anticipate the intuitive result (to be formalized in Eq. (3.44)) that the magnitude $|Z|$ of the complex impedance takes the place of the resistance R in Ohm's law. Some refinements will be needed to correctly manipulate these quantities in the complex plane, but in terms of magnitudes only, we may write $|V| = |I| |Z|$, using rms values for both voltage and current.

Thus, $|Z| = 120 \text{ V} / 8.33 \text{ A} = 14.4 \Omega$.

The power factor $\cos \theta$ corresponds to the ratio of resistance to impedance. Therefore, the resistive component is $R = 0.75 \cdot 14.4 \Omega = 10.8 \Omega$. The reactive component is proportional to $\sin \theta$ and equals $X = 0.661 \cdot 14.4 \Omega = 9.51 \Omega$.

We can express the impedance as the complex sum of its components: $\mathbf{Z} = 10.8 + j9.51 (\Omega)$. Alternatively, we can express it in terms of magnitude and angle: $\mathbf{Z} = 14.4 \angle 41.48^\circ \Omega$. The positive values of θ and X in this example indicate inductive (rather than capacitive) reactance.

With these numbers, we can verify that $Q = I^2 X = (8.33 \text{ A})^2 \cdot 9.51 \Omega = 660 \text{ VAR}$. However, we cannot apply the expression $Q = V^2 / X$ based on what we calculated above. This is because in writing $\mathbf{Z} = R + jX$ we implicitly modeled a resistance and reactance in series, where the voltage across X by itself does not actually equal 120 V.

Alternatively, the vacuum cleaner corresponding to the given P and Q could be modeled in terms of a parallel admittance $\mathbf{Y} = G + jB$. In that case, we would have $\mathbf{Y} = 1/\mathbf{Z} = 0.0694 \angle -41.48^\circ = 0.052 - j0.46 (\text{S})$. Since parallel elements share the same voltage, we could then write $P = V^2 G$ and $Q = V^2 B$, but we cannot use the total current. One lesson here is that working in the complex plane helps prevent mistakes that are easily made when trying to use scalars only.

3.4.4 Reactive Power in the Power System Context

The different power factors of individual loads combine into an aggregate power factor for the set of loads. The calculation is based directly on the combination of impedances or admittances in series and parallel. Although straightforward in principle, this calculation can become tedious in practice.

Qualitatively, we can say that if two loads of different power factors are connected in parallel, their combined power factor will be somewhere in the middle, and it will be closer to that of the bigger load. Just as an equivalent resistance can be defined for any combination of individual resistors, one can specify a power factor for a set of loads at any level of aggregation, from a single customer to a local area or the entire utility grid. Aggregate displacement power factors historically tended to be between unity and 0.9 lagging, but the proliferation of power electronic loads is changing this landscape (see Section 3.4.5).

Based on the law of energy conservation, the power going into a circuit must equal the power coming out of the circuit at every instant. In principle, therefore, inductive and capacitive behavior in a circuit must always be matched. Of course, this is not always guaranteed by design. For example, we may not know in advance the power factor of the load, which could vary in real-time. However, we generally expect that an a.c. power source behaves as a voltage source, and as such will be capable of making whatever adjustments necessary to provide current with the proper timing as demanded by the load.²⁸ If the source failed to do so in practice, it would be unable to maintain its voltage, and power flow in the circuit would collapse.

²⁸ How this works for traditional synchronous generators is detailed in Section 10.4.2, and for modern inverters in Section 14.4.

In operational terms, the problem of managing reactive power in the electric grid adds another dimension to managing real power: just like the system must collectively supply the precise number of watts demanded at any instant, it must arrange for the precise number of VARs conveyed at any instant. Because electric loads have been historically dominated by inductance, utilities have associated the provision of capacitive reactance with “producing” VARs, alongside real power.

Most customers, especially at the residential level, are only charged for the real power they consume.²⁹ Traditional analog kilowatt-hour meters can’t even measure reactive power, and typical residential smart meters don’t report it.

Even though “producing” reactive power does not inherently consume energy, providing and delivering it does have some costs for the utility. This is primarily because reactive power oscillating through the network requires additional current that produces unwanted heat. Transporting reactive power therefore entails some real energy losses, as well as greater capacity requirements—or an opportunity cost, if the available kVA capacity is tied up with VARs and thus able to carry fewer watts. This holds for transmission lines, conventional generators, and power electronics such as inverters. Owing to its property of occupying equipment while doing no useful work, reactive power has been jokingly called “the cholesterol of power lines.”³⁰

It is important to clarify some terminology about losses. The losses at issue here are real power losses in watts, associated with physical (I^2R) heating of power lines and other equipment, regardless of whether that current is attributable to real or reactive power drawn by the load (i.e., regardless of its timing relative to voltage). These I^2R line losses are not to be confused with the quantity called *reactive losses* in the industry. Reactive losses are an accounting term for the difference between VARs “supplied” by generators and VARs “demanded” by customer loads. The discrepancy is due to the inductive reactance of lines and transformers in the delivery system, which is neither directly observed nor controlled. If a transmission line has significant capacitance, reactive losses can be negative. This will be further discussed in the context of power flow analysis (Chapter 12). Although reactive losses don’t represent a direct cost to the system, they are important in the context of planning and scheduling reactive power provision.

Example

To illustrate the effect of the power factor on line losses, consider a load of 100 kW at the end of a several-mile-long 12-kV distribution line. Suppose the line’s resistance is 10 Ω. If the power factor is 0.8 lagging, the apparent power drawn by the load is 125 kVA, and the reactive power is 75 kVAR. The current to this load is $125 \text{ kVA} \div 12 \text{ kV} = 10.4 \text{ A}$.

The distribution line losses due to this load are given by $I^2R = (10.4 \text{ A})^2 \cdot 10 \Omega = 1.08 \text{ kW}$.³¹ This is significantly more than we might have expected just on the basis of real power demand: using only the real power of 100 kW, we would have estimated a current of 8.33 A and losses of only 0.69 kW.

3.4.5 Reactive Compensation

In order to minimize real I^2R losses and also maximize available equipment capacity, utilities take steps to bring the aggregate displacement power factor reasonably close to 1 (unity) with some form

²⁹ When large customers are charged for reactive power, it is often not per peak kVAR or cumulative kVAR-hour, but with a rate schedule for real power that is adjusted according to the customer’s power factor. For example, the monthly rate per kWh might increase if the power factor ever drops below 0.8.

³⁰ Another popular metaphor is foam on a beer, which fills a portion of the glass without delivering much substance. Unfortunately, these images offer no functional insight.

³¹ Realistically, there may be other loads along the same line. In that case, we would want to sum up the currents on each line segment before squaring to calculate losses.

of *reactive compensation*. Preferably, this compensation is placed near the load, so as to minimize the distance that the current associated with reactive power must travel through the infrastructure. In practice, because common loads tend to have lagging power factors, local reactive compensation involves capacitors.

Reactive compensation plays an even more important role for managing voltage levels across an a.c. network. Though it is crucial for operational stability, the relationship between reactive power and voltage is not at all intuitive. In essence, lagging currents tend to cause a larger voltage drop across transmission lines and other equipment, potentially causing unacceptably low voltages on the receiving end. Voltage regulation will be discussed further in Section 7.4.

With the proliferation of power electronics, especially motors with variable speed drives and LED lighting, typical load power factors are less lagging than they were historically. Besides induction motors, low power factor culprits included early-generation fluorescent lights with magnetic ballasts. Incandescent lamps and resistive heaters have power factors close to unity. The displacement power factor of a power electronic device, lagging or leading, can be anyone's guess.

More important for these devices is *harmonic distortion*, meaning that the load current is not sinusoidal (Section 5.3), even when presented with a sinusoidal voltage. Like displacement, harmonics imply current that is mismatched in time with the voltage, and consequently does no useful work—except that in the case of harmonics, the timing mismatch is an issue of frequency rather than phase shift. We will take up the associated *distortion power factor* in Section 5.3.3.

Absent harmonics, the amount of reactive compensation required to bring a given load to a desired power factor is straightforward to determine graphically, using the power triangle in the complex plane. If the original power factor is lagging, we introduce some parallel capacitance to “correct” it by simply adding the negative reactive power Q_C to offset Q_L from the load. The important insight is that real power remains unaffected. The process is best illustrated by example.

Example

Consider the load from the preceding example, with apparent power $S = 125 \text{ kVA}$ and real power $P = 100 \text{ kW}$, p.f. = 0.80 lagging. What size capacitor in kVA in parallel with this load would raise the combined power factor to 0.96 lagging?³²

The objective is to reduce the angle of lagging current from $\cos^{-1} 0.80 = 36.9^\circ$ to $\cos^{-1} 0.96 = 16.2^\circ$. At the new operating point, the combined reactive power demand will be $Q_{\text{TOT}} = S \sin 16.2^\circ = P \tan 16.2^\circ = 100 \text{ kW} \tan 16.2^\circ = 29 \text{ kVAR}$.

To find the required VAR contribution from the capacitor, we subtract the original demand of $Q_L = 125 \text{ kVA} \sin 36.9^\circ = 75 \text{ kVAR}$ to write $Q_C = Q_{\text{TOT}} - Q_L = -46 \text{ kVAR}$. The negative sign on Q_C indicates that “negative reactive power is being consumed”: in other words, 46 kVAR are being “produced” by the capacitor.

We might be interested in the reduction of total current to the load-capacitor combination. For this, we must solve explicitly for the new apparent power: $S_{\text{TOT}} = 100 \div \cos 16.2^\circ = 104 \text{ kVA}$. Then we can state that the current will be reduced by $(104 - 125) / 125 = 16.8\%$. Line losses, which depend on current squared, will be reduced by $1 - (1 - 0.168)^2 \approx 30\%$. This is an added benefit of installing the capacitor, besides managing voltage.

³² Why don't we choose to bring the power factor all the way to 1.0? First, we might want to save money and procure a smaller capacitor. More important, we may not want to oversize the fixed capacitor in case the load power factor will be higher at another time of day. Finally, 1.0 would give us much less opportunity to practice power factor arithmetic.

As in the above example, power factor correction capacitors are typically connected in parallel, known as *shunt* capacitance, rather than in series. One advantage of shunt capacitance is that its VAR contribution is a function of the line operating voltage ($Q = V^2/X$) and therefore more steady and predictable, rather than varying with load current in the series case ($Q = I^2X$). Connecting capacitors in parallel also avoids exposing them to potentially damaging fault currents, and makes it easy to switch them off rather than having to bypass them. Series capacitance is used in specific situations to compensate for transmission line reactance.

Example

If the 46-kVAR capacitor in the above example is connected at 12 kV and 60 Hz, what are its capacitive reactance in ohms, its admittance in siemens, and capacitance in farads?

The quick way to calculate reactance here is to apply $Q_C = V^2/X_C$, which yields $X_C = (12,000)^2 \div 46,000 = 3.13 \text{ k}\Omega$. Then $Y_C = 1/X_C = 0.32 \text{ mS}$ (milli-siemens) and $C = Y_C/60 \text{ Hz} = 5.3 \mu\text{F}$ (micro-farad).

The real capacitor nameplate would likely show the rating “46 kVAR, 12 kV, 60 Hz.” For purposes of circuit analysis, the utility engineer would describe capacitor reactance or admittance in the per-unit system (see Section 8.7).

3.5 Phasors

3.5.1 Introduction

Consider a voltage sine wave in the time domain. This is a physical quantity or signal that we can measure with an analog instrument. Now consider that there are other sine waves very much like it, that describe voltages or currents somewhere on the same circuit or network, that we would like to evaluate in relation to one another—for example, to apply Ohm’s law, or to compute power. We want to do this in the simplest, most straightforward way that captures the important information and does away with extraneous detail. To accomplish this, we take advantage of the fact that a sine wave is periodic: it’s perfectly predictable. If you’ve seen one cycle, you’ve seen them all.

For the purpose of our analysis, we make three important assumptions:

- (1) All these time-varying quantities are well described by a pure sinusoidal function of a single, fundamental frequency. In other words, there is no harmonic content, or distortion of the sinusoidal waveform.
- (2) The frequency does not vary in time.
- (3) The frequency does not vary by location; it is the same for voltages and currents everywhere in the circuit or network.

In other words, we assume the frequency of our sinusoidal functions is utterly uninteresting. Then we can avail ourselves of some clever mathematical tools to eliminate frequency from our expressions, and evaluate alternating voltages and currents strictly on the basis of two properties: the magnitude and the phase shift (or zero crossing) of each sine wave. If our three above assumptions hold, then magnitude and phase will tell us everything there is to know about the relationships of these quantities to each other, and they won’t ever change—at least not on a time scale relevant for our analysis.

It is important to emphasize that the above assumptions will limit us to a particular type of analysis: namely, an analysis of the *steady state* of an idealized network. We are obligated to remember

the limitations of our tools, if and when we encounter situations or states of the circuit we are analyzing that are no longer steady, nor well represented by pure sinusoidal functions. Such situations are increasingly common as electric grids evolve in the 21st century, and there is growing interest in physical measurements and analytic tools that could be better suited to them. Still, phasors are an indispensable tool for power system analysis, without which it would have been inconceivable to design and operate a.c. electric grids in the days of pencil and slide rule.

3.5.2 Derivation

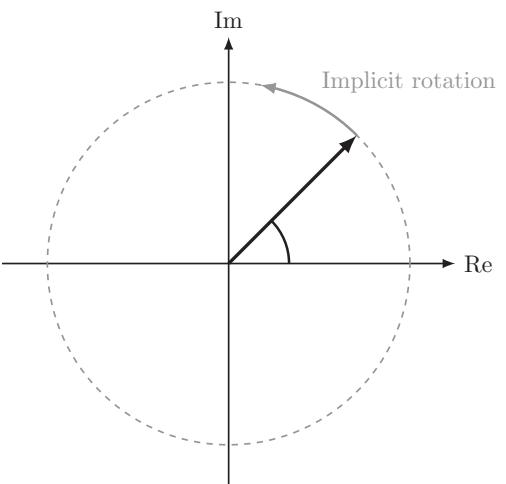
The familiar representation of a sine wave displays it with magnitude on the vertical axis and time on the horizontal axis. But, having asserted that the wave is periodic, there is no need to explicitly draw the time axis out into the future: we can be content with studying a single cycle. We can't get rid of time entirely, though, because we will want to depict relationships between and among zero-crossings of different waves. So we need to preserve one cycle's worth of time discrimination. To accomplish this, we employ a clever mathematical trick: namely, we introduce a third, imaginary dimension, to prepare our sine wave for better viewing. This will bring time shifts within a single cycle into focus, but ignore the repetition of multiple cycles.

To see how this works, we can imagine the sine wave as being the two-dimensional projection, or shadow, of a three-dimensional helix. Note that a cork screw, looked at from any side, appears exactly as a sinusoid like we draw on a graph. The dimension that is sticking out of the page, perpendicular to the plane on which the original sine wave was drawn, is called the imaginary axis. Movement along the helical path, when viewed from any side, is indistinguishable from an up-down oscillating motion. That one-dimensional projection, onto the real axis, is the only thing we can physically measure. But there is no harm in conceiving of it as the result of a more complex helical motion; no information is lost or distorted.

Now, instead of looking at the wave from the side, let's look at the helix head-on. From here, it looks like a circle! That circle can be depicted in the complex plane. We imagine the time-varying voltage as a point that travels around the circumference at a constant speed, or as an arrow—which we will call a *phasor*—pointing to it from the origin. This is shown in Figure 3.14.

Again, the oscillation of the physically measurable voltage corresponds to the projection of that point or arrow onto the real axis. The imaginary direction doesn't correspond to anything physical,

Figure 3.14 A phasor in the complex plane.



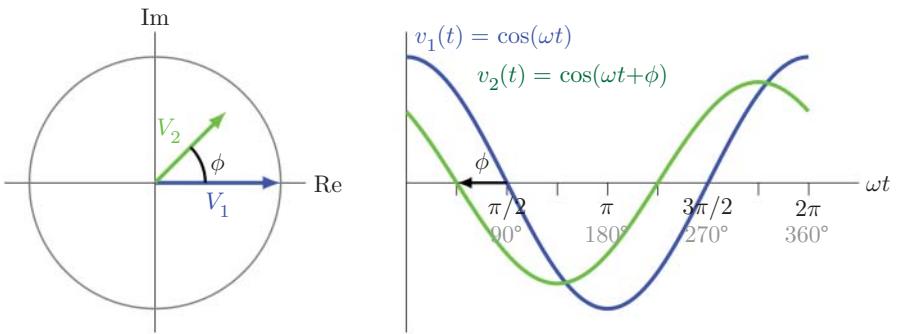


Figure 3.15 Phasors in relation to each other.

but it also doesn't invalidate anything. Rather, the complex plane gives us a space for drawing out temporal relationships without making explicit reference to time.

For example, consider two sinusoidal voltages, one lagging the other by 45° . In other words, its zero crossing always occurs one-eighth of a cycle later than the first. We can draw the two phasors relative to each other with lengths proportional to their magnitudes, at an angle of 45° to each other. Both of them are imagined as rotating at a constant speed, in constant relation to each other. Figure 3.15 illustrates the situation, where we have chosen V_1 to have an angle of zero.

Note that for the sake of capturing the relationship between the two phasors, it doesn't matter which coordinate system we choose (i.e., when to call the time $t = 0$), as long as the same is used for both. Regardless of when we start counting time, we will always observe the same 45° separation between the two.

We do, however, need to decide which direction they turn: clockwise or counter-clockwise. There is no physical meaning to this choice; it is just a convention for internal consistency. To be consistent with the definition of mathematical operations in the complex plane, we let all phasors rotate counterclockwise in time.

We can draw any number of phasors into the same diagram, both voltages and currents. One thing to note is that since current and voltage are measured in different physical units, their comparative scale on the graph (or length of the arrows) is arbitrary. This does not affect the timing (or angle) between them.

Let us now translate the above visualization of the helix into mathematical notation. Expanding the sine wave into the imaginary direction amounts to writing the familiar sinusoid as the real part of a complex expression:

$$v(t) = \operatorname{Re}\{\mathbf{v}(t)\} \text{ where } \mathbf{v}(t) = \cos(\omega t + \phi) + j \sin(\omega t + \phi) \quad (3.35)$$

We have effectively expanded one sinusoidal function into two, one of which is purely imaginary. The real and imaginary parts of this compound oscillation are offset by 90° from each other.³³

Now we will use Euler's equation³⁴ to represent sinusoidal functions as exponential expressions:

$$e^{jx} = \cos x + j \sin x \quad (3.36)$$

³³ Note that if we had chosen sine for the real part and cosine for the imaginary part, it would make quantities rotate clockwise instead of counterclockwise, and not be consistent with standard convention.

³⁴ For readers not familiar with Euler's equation, it is worth contemplating in the optional sidebar below.

We substitute $x = (\omega t + \phi)$. Because exponents add when we multiply terms, we can expand this into the product of two exponentials, one of which is a function of time and one of which isn't:

$$v(t) = \operatorname{Re}\{V_{\max} e^{j(\omega t + \phi)}\} = \operatorname{Re}\{V_{\max} e^{j\omega t} e^{j\phi}\} \quad (3.37)$$

The term $e^{j\omega t}$ is sometimes called the *rotating phasor*, as it simply spins in circles, and $e^{j\phi}$ the *stationary phasor*, which contains information about the relative phase or starting point compared to other sinusoids of the same frequency.

Now comes the moment for the draconian step: throwing out the time dependence and frequency. On the assumption that $e^{j\omega t}$ contains nothing interesting, we simply omit the rotating phasor and abbreviate the expression. This transforms the time-varying $v(t)$ into a static complex number that preserves only two pieces of information, amplitude and phase angle:

$$v(t) \Rightarrow \mathbf{V} = V_{\max} e^{j\phi} \quad (3.38)$$

This complex number, which technically retains only the *stationary phasor*, is simply called the *phasor* by almost every power engineer. We will further address this all-important simplification and its associated caveats in Section 16.2.4.

We are almost finished, but not quite. In power engineering, a convention is preferred where instead of the amplitude, we use the root-mean-square (rms) magnitude for the phasor. The reason is that we will often multiply voltage and current phasors together to obtain power, where the rms convention will save us the trouble of including a factor of $\frac{1}{2}$. Finally, we will abbreviate the exponential to show only the informative part, ϕ , preceded by the angle symbol. Thus, the standard power engineering phasor, as used in the rest of this book, is written:

$$v(t) = V_{\max} \cos(\omega t + \phi) \Rightarrow \mathbf{V} = V_{\text{rms}} e^{j\phi} = V_{\text{rms}} \angle \phi \quad (3.39)$$

Current and voltage phasors are completely analogous.

3.5.3 Euler's Equation

What does an exponential function have to do with sine waves? Not much, until we put imaginary numbers in the exponent. In the real world, we are accustomed to exponential functions doing either of two things: grow (if the exponent is increasing, as in e^x), or decay (if the exponent is decreasing, as in e^{-x}). A complex exponential instead rotates in the complex plane. Just as multiplying a number by j corresponds to performing an operation that rotates it counterclockwise in the complex plane by 90° , raising a number to an exponent with some multiple of j means rotating it counterclockwise by some amount (or clockwise for $-j$), as illustrated in Figure 3.16. The real part of a complex exponent will grow or shrink the number, and the imaginary part will only rotate it.

Regular human intuition fails us here because, even worse than non-integer exponents, complex exponents don't translate at all into the notion of multiplying the base by itself some number of times. So we must resign ourselves to trusting the mathematical rules of operation—like flying an aircraft by instruments only. Following rules of operation, any real number N raised to a complex exponent $C = a + jb$ can be written as

$$N^C = N^{a+jb} = N^a \cdot N^{jb}$$

where N^a scales the magnitude of the result, and N^{jb} has a magnitude of 1 and only rotates in angle. Depending on the base N , as b increases, the value of N^{jb} revolves faster or slower around the unit circle in the complex plane.

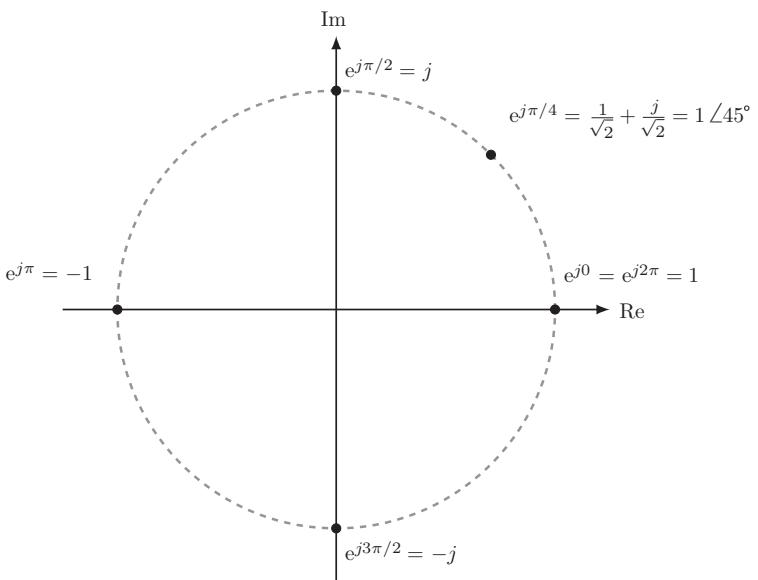


Figure 3.16 Complex exponentials. An increasing imaginary exponent corresponds to counterclockwise rotation in the complex plane.

The number $e = 2.71828\dots$, known as the base of the natural logarithm, plays a special role here. It is the base for which e^{ix} completes exactly a full circle when $x = 2\pi$. Stated another way, the angular rotation of e^{ix} in the complex plane is always exactly equal to x in radians.

The value of e can be formally represented in terms of various infinite series expansions.³⁵ The expression e^x can be written as the infinite series:

$$e^x = 1 + \frac{x}{1!} + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \dots \quad (3.40)$$

Although e^x grows rapidly for increasing x , the series always converges to a finite value because the factorial $n! = 1 \cdot 2 \cdot 3 \cdots n$ eventually grows faster than x^n .

Recalling the rules for differentiating polynomials from introductory calculus, notice what happens when we take the derivative of e^x : each term in the series turns into the previous term. The value of the entire series is thus unchanged by differentiation with respect to x , and e^x is famously the only function that is always identical to its own rate of change. We can define e as the unique number for which that statement is true.

Readers steeped in trigonometric functions may recall that $\sin(x)$ and $\cos(x)$ can be written similarly, but with alternating $+/ -$ signs, and with sine getting only the odd and cosine the even terms:

$$\sin(x) = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \frac{x^9}{9!} - \dots \quad (3.41)$$

$$\cos(x) = 1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \frac{x^8}{8!} - \dots \quad (3.42)$$

Notice that while e^x either grows or shrinks with x , the values of $\sin(x)$ and $\cos(x)$ oscillate with growing x , as alternately the positive or negative terms in the series dominate.

³⁵ The Wikipedia article for *e (mathematical constant)* offers clear definitions and interesting context. For a detailed historical account, see Eli Maor, *e: The Story of a Number* (Princeton Science Library, 1994). I can't resist pointing out that this should be available as an e-book.

Now we can appreciate the momentous significance of introducing the j in the exponent. Because j^n rotates through positive and negative, imaginary and real versions of unity, it produces the alternation we see in the trig functions:

$$e^{jx} = 1 + \frac{jx}{1!} + \frac{-x^2}{2!} + \frac{-jx^3}{3!} + \frac{x^4}{4!} + \frac{jx^5}{5!} + \frac{-x^6}{6!} + \frac{-jx^7}{7!} + \frac{x^8}{8!} + \dots \quad (3.43)$$

We only need to group the terms with and without the j , and we get back $\cos(x) + j \sin(x)$. Leonhard Euler figured this out in the 1740s, but it's never too late to be amazed.

3.5.4 Operations with Phasors

Understanding phasors is hard; using them is easy. Since we have effectively converted time-varying quantities into static complex numbers, all we need to do is apply the rules of complex arithmetic for performing operations on them. Sinusoidal voltages and currents, and even power, can now be manipulated just like constant impedances and admittances. Doing this graphically is often easier than numerically.

Any common operation will be based on one of the fundamental relationships in electric circuits: Ohm's law, Kirchhoff's voltage or current law, and power as the product of voltage and current.

3.5.5 Ohm's Law in Complex Form

In phasor notation, we write³⁶ the complex form of Ohm's law in terms of impedance as

$$\mathbf{V} = \mathbf{I}\mathbf{Z} \quad \text{or} \quad V\angle\phi_V = I\angle\phi_I \cdot Z\angle\theta \quad (3.44)$$

If we choose voltage as the reference and set $\angle\phi_V = 0$, then the rules of multiplying complex numbers (adding their angles) require that the current angle is negative that of the impedance, $\angle\phi_I = -\theta$.

Conversely, we write Ohm's law in terms of admittance as

$$\mathbf{I} = \mathbf{VY} \quad \text{or} \quad I\angle\phi_I = V\angle\phi_V \cdot Y\angle-\theta \quad (3.45)$$

Letting $\angle\phi_V = 0$ we see that the current angle must match the admittance, $\angle\phi_I = -\theta$.

Example

Consider a load with total series impedance $\mathbf{Z}_s = 1.8\angle56.3^\circ$ (as depicted in an earlier example), supplied by a sinusoidal voltage source $v(t) = 169 \cos(377t + 10.0^\circ)$. In phasor notation, the voltage becomes $\mathbf{V} = 120\angle10.0^\circ \text{ V}$.

To find the current, we first write the admittance:

$$\mathbf{Y}_s = 1/\mathbf{Z}_s = 0.555\angle-56.3^\circ \text{ S.}$$

By Ohm's law, the current is

$$\mathbf{I}_s = \mathbf{VY}_s = 120\angle10.0^\circ \text{ V} \cdot 0.555\angle-56.3^\circ \text{ S} = 66.7\angle(10 - 56.3)^\circ \text{ A} = 66.7\angle-46.3 \text{ A.}$$

Note that the explicit assignment of $\angle I_s$ is determined by the voltage reference $v(t)$ (specifically, the voltage phase angle at $t = 0$), although the current will always lag the voltage by 56.3° regardless of the choice of coordinate system.

³⁶ Here we emphasize complex quantities with **boldface** notation, to distinguish them from scalars (plain numbers). In later chapters, where it is assumed obvious from context what is a complex quantity and what is a scalar, we will relax this conscientious practice.

3.5.6 Kirchhoff's Laws with Phasors

To add voltage drops across individual series elements around a circuit, we perform vector addition with phasors. Similarly, we add current phasors at a circuit node. The process is best illustrated by example, using a series and parallel circuit, respectively.

Example

Let's reuse the series circuit from Figure 3.10 again, now choosing a convenient coordinate system where $\mathbf{V} = 120\angle 0^\circ$ V. The combined impedance is $\mathbf{Z}_s = 1.8\angle 56.3^\circ \Omega$, the combined admittance is $\mathbf{Y}_s = 0.555\angle -56.3^\circ \text{ S}$, and the current phasor is $\mathbf{I}_s = \mathbf{V}\mathbf{Y}_s = 66.67\angle -56.3^\circ \text{ A}$. The key point to recognize is that the current \mathbf{I}_s through each series element must be the same.

Nevertheless, each individual circuit element prescribes a different temporal relationship between voltage and current. How is this possible? Because not every element "sees" the entire voltage source. While the voltages across each series element must, by KVL, add up to the source voltage $v(t)$ at every instant, their magnitudes and their phases are shifted so as to satisfy Ohm's law individually as well as collectively.

To illustrate, we label points a, b, c, d on the circuit diagram in Figure 3.17. The voltage drop across the resistor, between a and b, is given by

$$\mathbf{V}_R = \mathbf{I}\mathbf{R} = 66.7\angle -56.3^\circ \text{ A} \cdot 1\Omega = 66.7\angle -56.3^\circ \text{ V}$$

The voltage drop across the inductor (whose impedance $j2\Omega$ we transcribe as $2\angle 90^\circ$), between points b and c, is

$$\mathbf{V}_L = \mathbf{I}\mathbf{X}_L = 66.7\angle -56.3^\circ \text{ A} \cdot 2\angle 90^\circ \Omega = 133\angle 33.7^\circ \text{ V}$$

For the capacitor³⁷ between points c and d, we have

$$\mathbf{V}_C = \mathbf{I}\mathbf{X}_C = 66.7\angle -56.3^\circ \text{ A} \cdot 0.5\angle -90^\circ \Omega = 33.35\angle -146.3^\circ \text{ V}$$

Thus, each of the three circuit elements experiences a distinct voltage drop, in both magnitude and timing. The fact that these three voltages *always* add up to the source voltage is stated concisely by the phasor addition

$$\mathbf{V} = \mathbf{V}_R + \mathbf{V}_L + \mathbf{V}_C$$

without any explicit reference to time. Equivalently, we could write KVL in the time domain, for all t :

$$v_{a-b}(t) + v_{b-c}(t) + v_{c-d}(t) + v_{d-a}(t) = 0$$

Though the phasor addition can be shown arithmetically, it is best seen graphically, as sketched in Figure 3.18 (approximately to scale). The entire phasor diagram can be visualized as rotating

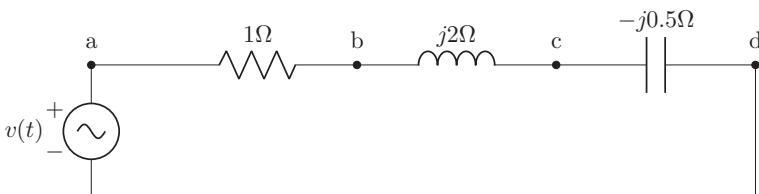


Figure 3.17 Series circuit to illustrate KVL for the complex case.

³⁷ Note that it doesn't matter how we transcribe $-j0.5$: If instead of $0.5\angle -90^\circ$ we write $-0.5\angle 90^\circ$, our solution appears as $V_C = -33.35\angle 34.44^\circ$, which describes the same vector in the complex plane.

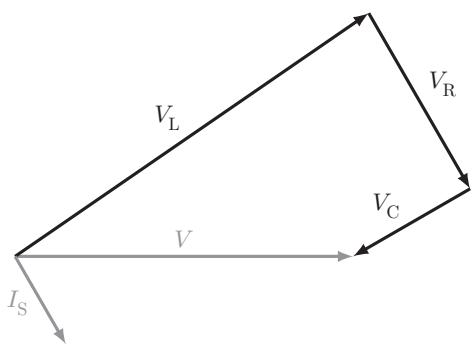


Figure 3.18 Phasor addition of voltages $V = V_R + V_L + V_C$ associated with Figure 3.17 in the complex plane. Note that the length of the current phasor is arbitrary since it is in different units; it is shown for angle reference only.

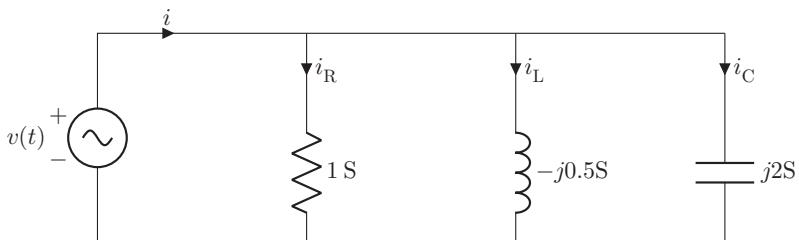


Figure 3.19 Parallel circuit to illustrate KCL for the complex case.

counterclockwise at the synchronous frequency ω . Choosing a different phase angle for $v(t)$, here drawn horizontally corresponding to $\angle V = 0^\circ$, would simply amount to taking a snapshot of the entire diagram in a different orientation, with the relationship among all the phasors preserved.

Example

The analogous addition of current phasors is illustrated with the corresponding parallel circuit, reproduced in Figure 3.19 to highlight the branch currents, and phasor diagram shown in Figure 3.20.

We now observe that the voltage drop across each parallel element is the same, but the branch currents will differ in both magnitude and phase. Choosing the same voltage as in the series circuit, we write for the resistor $\mathbf{V} = \mathbf{I}_R R$, or

$$\mathbf{I}_R = \mathbf{V} \mathbf{Y}_R = \mathbf{V} G = 120 \angle 0^\circ \text{V} \cdot 1 \angle 0^\circ \text{S} = 120 \angle 0^\circ \text{A}$$

For the ideal inductor, $\mathbf{V} = \mathbf{I}_L X_L$, or

$$\mathbf{I}_L = \mathbf{V} \mathbf{Y}_L = \mathbf{V} (-jB_L) = 120 \angle 0^\circ \text{V} \cdot 0.5 \angle -90^\circ \text{S} = 60 \angle -90^\circ \text{A}$$

For the ideal capacitor, $\mathbf{V} = \mathbf{I}_C X_C$, or

$$\mathbf{I}_C = \mathbf{V} \mathbf{Y}_C = \mathbf{V} B_C = 120 \angle 0^\circ \text{V} \cdot 2 \angle 90^\circ \Omega = 240 \angle 90^\circ \text{A}$$

where the branch current has twice the magnitude as for the resistor.

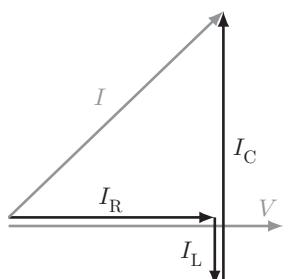


Figure 3.20 Phasor addition in the complex plane of currents
 $I = I_R + I_L + I_C$ associated with Figure 3.19.

KCL requires that the currents add to zero at the common nodes, which correspond to the entire upper and lower edge of the circuit diagram.³⁸ Thus, the total current equals the sum of branch currents, written in phasors as

$$\mathbf{I} = \mathbf{I}_R + \mathbf{I}_L + \mathbf{I}_C$$

or in the time domain as

$$i(t) = i_R(t) + i_L(t) + i_C(t)$$

As illustrated in the phasor diagram in Figure 3.20, the sum of the three complex branch currents is

$$\mathbf{I} = 120 - j60 + j240 = 120 + j180 = 216.3\angle 56.3^\circ \text{ A}$$

which is consistent with applying $\mathbf{I} = \mathbf{V}\mathbf{Y}$ to the parallel combination. Notice that the parallel combination produces power factor $\cos -56.3^\circ = 0.55$ leading, while the series combination of the same three elements had $p.f. = \cos(56.3^\circ) = 0.55$ lagging.

3.5.7 Complex Power in Phasor Notation

Finally, we write complex power in terms of voltage and current phasors:

$$\mathbf{S} = \mathbf{V}\mathbf{I}^* \quad (3.46)$$

If this entire book were to be summarized in a single equation, it would be 3.46. We already know power is the product of voltage and current, but perhaps more impressive is that the phasor notation accommodates both real and reactive power in such a concise expression. As noted earlier, the complex conjugate on the current means that we reverse the sign of the imaginary component, or the angle. This ensures that we take the *difference* between voltage and current phase angles, not their sum. This difference is a physically meaningful quantity, reflecting the phase shift demanded by the impedance.

Recall from Eq. (3.27) for instantaneous power that

$$p(t) = v(t)i(t) = V_{\text{rms}}I_{\text{rms}} \cos(\phi_V - \phi_I) + V_{\text{rms}}I_{\text{rms}} \cos(2\omega t + \phi_V + \phi_I) \quad (3.47)$$

Writing this as the real part of a complex exponential, we get

$$p(t) = V_{\text{rms}}I_{\text{rms}} \operatorname{Re}\{e^{j(\phi_V - \phi_I)}\} + V_{\text{rms}}I_{\text{rms}} \operatorname{Re}\{e^{j(2\omega t + \phi_V + \phi_I)}\} \quad (3.48)$$

Next, we discard the time-varying second term to write complex power as a steady-state quantity that can be mapped statically in the complex plane.

$$p(t) \Rightarrow \mathbf{S} = V_{\text{rms}}I_{\text{rms}} e^{j(\phi_V - \phi_I)} = S\angle(\phi_V - \phi_I) \quad (3.49)$$

Despite the frequency of power being twice that of voltage and current, we can still throw out the explicit time variation and retain an eminently useful quantity. The discrepancy between the frequencies of \mathbf{V} , \mathbf{I} , and \mathbf{S} means their relationship will vary over the course of the cycle. Unlike the phasor sum diagram for voltages and currents, which we can visualize as rotating all in one piece, a diagram attempting to illustrate the product of voltage and current would hardly be meaningful, as the product would move around relative to \mathbf{V} and \mathbf{I} and make two revolutions for every one cycle.

³⁸ The “node” can take any shape but remains the same point electrically, since the connecting “wires” in our diagram are assumed to be ideal without any impedance whatsoever.

Nevertheless, the magnitude S remains constant, and its waxing and waning projection onto the real axis correspond to the power delivered to the load.

Example

Consider two ideal voltage sources connected by a transmission line, as shown in Figure 3.21. We might imagine them as rotating machines or inverters, but they are abstracted here as one-ports (Section 2.5.1). Either machine could be generating or consuming real and reactive power in any combination.³⁹ Without knowing what's inside, we assume that each source has whatever physical resources necessary to maintain the stated voltage under any condition, regardless of the current and its relative timing. By convention, if positive current flowing *out* of the machine coincides with positive voltage, the machine is generating power; if positive current flows *into* the machine with positive voltage, it is acting as a load and consuming power.

Suppose the source voltages are $\mathbf{V}_1 = 100\angle 0^\circ$ V and $\mathbf{V}_2 = 100\angle 30^\circ$ V, respectively, and the impedance connecting them is purely inductive at $Z = 0 + j5\Omega$. Which machine is generating, and which is absorbing power? What is the balance of reactive power, and what is the role of the transmission line?

We begin by choosing a positive reference direction for the current, from Machine 1 to Machine 2. The current is given by

$$\begin{aligned}\mathbf{I} &= \frac{\mathbf{V}_1 - \mathbf{V}_2}{Z} = \frac{100 + j0 - (86.6 + j50)}{j5} \\ &= \frac{13.4 - j50}{j5} = -10 - j2.68 = 10.35\angle 195^\circ \text{ A}\end{aligned}$$

Since \mathbf{I} is defined as *leaving* Machine 1, the complex power $\mathbf{S} = \mathbf{VI}^*$ represents its generation, or injection into the circuit. For symmetry, let's express the power *entering* each machine, which means that we'll reverse the sign of current for Machine 1:

$$\begin{aligned}\mathbf{S}_1 &= \mathbf{V}_1(-\mathbf{I})^* = 100\angle 0^\circ \cdot -10.35\angle -195^\circ = -1035\angle -195^\circ = 1035\angle -15^\circ \text{ VA} \\ &= P_1 + jQ_1 = 1000 \text{ W} - j268 \text{ VAR}\end{aligned}$$

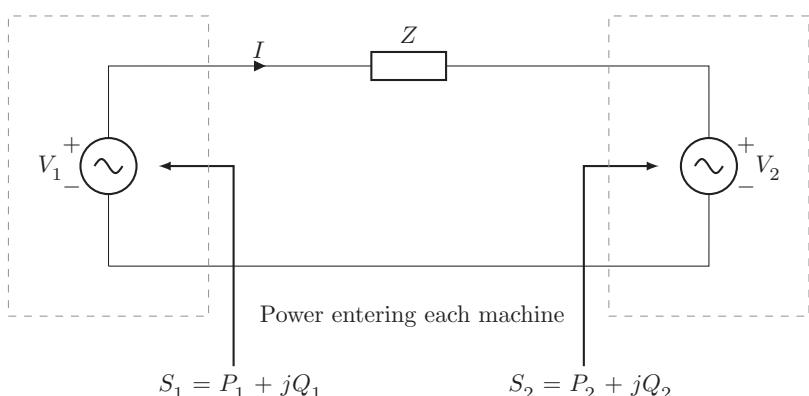


Figure 3.21 Two voltage sources that could be acting as generators or loads.

³⁹ This is known in practice as “four quadrant” operation.

The positive result for P_1 means that this machine is acting as a load, dissipating real power. The negative result for Q_1 in the load convention means that this machine is acting as a VAR source, “supplying” reactive power like a capacitive load. For Machine 2, we retain the positive reference direction of current into the machine (load convention):

$$\begin{aligned}\mathbf{S}_2 &= \mathbf{V}_2 \mathbf{I}^* = 100\angle 30^\circ \times 10.35\angle -195^\circ = 1035\angle -165^\circ \text{ VA} \\ &= P_2 + jQ_2 = -1000 \text{ W} - j268 \text{ VAR}\end{aligned}$$

The negative result for P_2 means that this machine is acting as a generator, providing real power to the circuit. Like Machine 1, its negative reactive load means that it is acting as a VAR source (in an equal amount).

Since the transmission line has zero resistance, it should dissipate no real power. This is consistent with our finding that Machine 2 generates exactly as much real power as Machine 1 consumes. As for reactive power, we would expect the transmission line to “consume” VARs, since it is inductive. Absent any other circuit elements in the picture, the line should account for the total amount of reactive power generated by the two machines, $268 + 268 = 536$ VAR. We can confirm this result by calculating the *reactive losses* as

$$Q_L = I^2 X = 10.35^2 \cdot 5 = 536 \text{ VAR}$$

Beyond the sign convention, this example illustrates two important observations about transferring power across transmission lines, which in practice are most commonly dominated by inductive properties. First, all else being equal, real power flows from the greater to the lesser voltage phase angle. In the example, the fact that the voltage phase angle for \mathbf{V}_2 is 30° greater than for \mathbf{V}_1 tells us that the source at \mathbf{V}_2 is injecting P . Second, reactive power injection is closely associated with voltage magnitude. Since the voltage magnitudes in the example are equal, the two machines share equally in their contribution to reactive power in the system.

Problems and Questions

- 3.1** Write the following complex numbers in polar coordinates: $C = 5 + j2$, $D = -5 + j2$. Determine the following: $-C$, $-D$, C^* , D^* , $1/C$, $1/D$, $C + D$, $C - D$, $C \cdot D$, C/D , D/C .
- 3.2** Are complex numbers a discovery or an invention? Discuss.
- 3.3** State the following as complex numbers in polar and rectangular coordinates: $e^{j\pi/6}$, $e^{j\pi/4}$, $e^{j3\pi/4}$, $e^{j5\pi/4}$, $e^{j7\pi/4}$.
- 3.4** The impedance Z of an electrical device is $Z = 3 + j4 \Omega$.
 - Write Z in polar coordinates, with the angle in degrees.
 - Find the admittance Y in polar coordinates.
 - Find the conductance G and the susceptance B .
 - Does the ratio $|G| / |B|$ match your expectations? Explain.
- 3.5** A 12-V, 50-Hz sinusoidal source is connected to a series combination of a $3-\Omega$ resistance, an $8-\Omega$ inductance, and a $0.25-\text{S}$ capacitance.
 - Draw a circuit diagram.
 - Find the series impedance at 50 Hz.

- (c) Find the series admittance at 50 Hz.
 (d) Find the current delivered by the source. Is it lagging or leading?
 (e) Find the real power dissipated and the (displacement) power factor of this circuit.
- 3.6** Repeat the previous problem, with the same three devices connected in parallel. Discuss how and why the answer to part (e) is different.
- 3.7** A mile-long conductor has a series inductive reactance of $X_L = j0.4 \Omega$ for an a.c. frequency of 60 Hz.
 (a) What is its inductance in henrys?
 (b) What would be its inductive reactance at 400 Hz?
 (c) What parallel capacitance in farads would create an equivalent admittance (susceptance) of $Y_C = j2.5 \text{ S}$ at 60 Hz?
 (d) What parallel capacitance would create an equivalent admittance at 400 Hz?
- 3.8** By convention, power factors of inductive and capacitive loads are described as “lagging” and “leading,” respectively. An alternative convention uses *sign* to distinguish the two cases, defining a lagging power factor as positive and a leading power factor as negative. What do you think are the advantages and drawbacks of that usage, and which convention would you recommend as being least likely to confuse people?
- 3.9** Comment on the distinction between “producing” and “consuming” reactive power. What is problematic about those words? Could you suggest a better terminology?
- 3.10** A vacuum cleaner with an induction motor is rated for operating on 120 V and is labeled “12.0A.” Based on this information, what can you say about its power consumption? Can you calculate how many kWh of energy it will use during a one-hour vacuuming session? Explain.
- 3.11** An a.c. voltage of 240 V is applied to a series circuit whose impedance is $10\angle60^\circ \Omega$.
 (a) Find R , X , S , P , Q , and the (displacement) power factor of the circuit.
 (b) If a capacitor is connected in parallel with this circuit and supplies 1250 VAR, find P and Q supplied by the 240-V source, and the resultant power factor.
- 3.12** A circuit supplied by a 120-V a.c. source has a resistance of 8Ω in series with an inductive reactance of $j6\Omega$.
 (a) What is the current phasor?
 (b) What are S , P , and Q ?
 (c) Suppose you wish to improve the power factor of the entire circuit to 0.95 lagging, by placing a capacitor in parallel with the $R-X_L$ combination. What is the value of the capacitive reactance X_C required?
 (d) What is the capacitance in farads, at 50 Hz and at 60 Hz?
- 3.13** In the power factor correction scenario from the previous problem, how many joules of energy are exchanged every half-cycle between the inductor and the capacitor, and how many joules between the inductor and the voltage source, at 50 Hz and at 60 Hz?

- 3.14** What assumptions are made in using the phasor representation of a.c. quantities? Describe an example of a situation that would be poorly represented by the phasors introduced in this chapter.
- 3.15** A phasor measurement unit (PMU) is an instrument that can determine the phase shift of an a.c. voltage or current relative to a reference clock, or relative to another voltage or current. Synchronized timekeeping (for example, using a pulse-per-second signal from GPS satellites) is essential for PMUs. Suppose you wish to resolve a phase shift between two voltages measured at different locations, to within 0.1° of angle (which determines the required resolution of the common timing source). How long, in microseconds, is a 0.1° time interval at 50 Hz, and at 60 Hz?
- 3.16** The accuracy of PMUs has sometimes been described in terms of a “Total Vector Error” (TVE). An early performance standard called for an accuracy of 1% TVE, meaning that acceptable measurements lie within a circle of radius 0.01 times the magnitude, in the complex plane. This error could come from some combination of magnitude or angle error, as long as the tip of the measured vector lies within the circle. For example, if the angle measurement is perfect, then a ± 0.01 error in the magnitude measurement is acceptable. Calculate the acceptable angle error in degrees by this standard, for the case of a perfect magnitude measurement.
- 3.17** Let $v(t) = 141.4 \cos(\omega t - 60^\circ)$ V and $i(t) = 11.31 \cos(\omega t - 30^\circ)$ A.
- What are the maximum value and rms value for each?
 - Draw the two waveforms in the time domain.
 - State a phasor expression for each in both polar and rectangular form, and draw a phasor diagram showing voltage and current in the complex plane.
 - Change your coordinate system such that the voltage angle is zero, and restate the phasors in polar form. Does this change the phasor diagram?
 - Is this circuit inductive or capacitive?
- 3.18** A voltage $V = 100\angle 0^\circ$ V and a current $I = 8\angle 60^\circ$ A are measured at a one-port.
- Suppose you know this circuit contains a resistive and a reactive element in series. Find the values of R_s and X_s .
 - Suppose instead that you know the resistive and reactive element are connected in parallel. What are the values of R_p and X_p for this case?
 - Explain in your own words why the comparative magnitudes of R_s , X_s , R_p , and X_p make sense.
- 3.19** A motor load is plugged into a wall outlet at the far end of the house, away from the utility service panel, through a stretch of wiring with a series impedance of $0.3 + j0.4\Omega$. The outlet voltage at no load is 120 V. The motor momentarily draws a starting current of 60 A.
- Suppose the starting current were in phase with the source voltage. What would be the voltage magnitude measured at the outlet during that disturbance?
 - Now suppose (more realistically) the starting current is lagging the source voltage by 70° . What is the voltage magnitude measured at the outlet in this case?
 - Do you think you would notice lights on the same circuit dimming?

- 3.20** Consider the situation illustrated in Figure 3.21, where two voltage sources are connected by a purely inductive line with impedance $Z = jX$. The voltage sources could represent generators or loads; all we know about them is that they maintain a certain voltage.
- (a) Write down general expressions for P_1 , Q_1 , P_2 , and Q_2 as the real and reactive power injected by Source 1 and Source 2, in terms of the voltages at either end of the transmission line and the line impedance. How much power is being transferred from one source to the other, and how much power is transferred to the line?
- (b) Derive the condition for maximum real power transfer between Source 1 and 2 to occur.
- 3.21** Referring again to the situation from Figure 3.21, change the voltage and impedance values according to the scenarios below (in volts and ohms). Determine the current and the real and reactive power injected or consumed by each source and by the transmission line, and summarize the energy balance and general intuition for each scenario in a sentence. Draw a phasor diagram for each scenario, indicating V_1 , V_2 , and the voltage drop IZ across the transmission line. Note that not all of these scenarios are physically realistic.
- (a) $V_1 = 20\angle 30^\circ$, $V_2 = 20\angle 0^\circ$, and $Z = 0 + j1$.
- (b) $V_1 = 20\angle 45^\circ$, $V_2 = 20\angle 0^\circ$, and $Z = 0 + j1$.
- (c) $V_1 = 20\angle 30^\circ$, $V_2 = 20\angle 0^\circ$, and $Z = 0 - j1$.
- (d) $V_1 = 20\angle 0^\circ$, $V_2 = 10\angle 0^\circ$, and $Z = 0 + j1$.
- (e) $V_1 = 20\angle 30^\circ$, $V_2 = 20\angle 0^\circ$, and $Z = 1 + j0$.
- (f) $V_1 = 20\angle 30^\circ$, $V_2 = 20\angle 0^\circ$, and $Z = 1 + j1$.

4

Three-Phase Power

4.1 Three-Phase Basics

In Chapter 3, we considered alternating-current (a.c.) circuits with a single phase, which is sufficient to introduce all the basic analytic tools. In reality, standard circuits in large power systems generally include three electrically separate parts or phases. For example, transmission lines are typically seen with sets of three conductors. Three-phase circuits can often be analyzed on a per-phase basis and represented by a single line in a *one-line diagram*. Still, it is sometimes necessary to work with three phases explicitly. This section describes the three-phase concept—how it works, why it was chosen, and basic techniques for analyzing three-phase circuits.

4.1.1 Rationale for Three Phases

There are essentially two reasons for configuring alternating current in three phases: economy of transmission, and efficiency of power conversion in rotating machines.

Three-phase transmission permits the use of fewer wires than common sense might suggest, and indeed less conductor capacity than would be required to transmit an equivalent amount of power using only a single phase of a.c. In general, the completion of an electric circuit requires two conductors between the power source and the load: one for the current to flow out and one for it to return.¹ This requirement is finessed by using multiple phases that are staggered in time.

We begin by imagining power being supplied from a source to a load in a single a.c. phase. Two conductors between the source and the load form one complete circuit. At any instant, the current everywhere along this circuit is the same. (This implies that if the conductors to and from the load are placed side by side, their currents go in opposite directions.)

Let us now supply three separate circuits, each with its own voltage source—that is, a set of generator windings—and set of loads. Each of these circuits will again have two conductors, one leading from one end of the generator armature winding to the load, and one connecting back from the load to the other end of the winding. Thus, we have a total of six wires.

The secret of multiphase transmission is that if we cleverly stagger the timing of a.c. on each of the circuits, they can serve as return conductors for one another. Specifically, for three phases, we

¹ It is sometimes possible to do away with one of the conductors and allow the current to return through the ground, but this is usually impractical. In most places, the soil presents too high an impedance for alternating current. High-voltage d.c. transmission is more likely able to operate with ground return, but this is rarely preferred. In any case, multi-phase transmission does not rely on ground return.

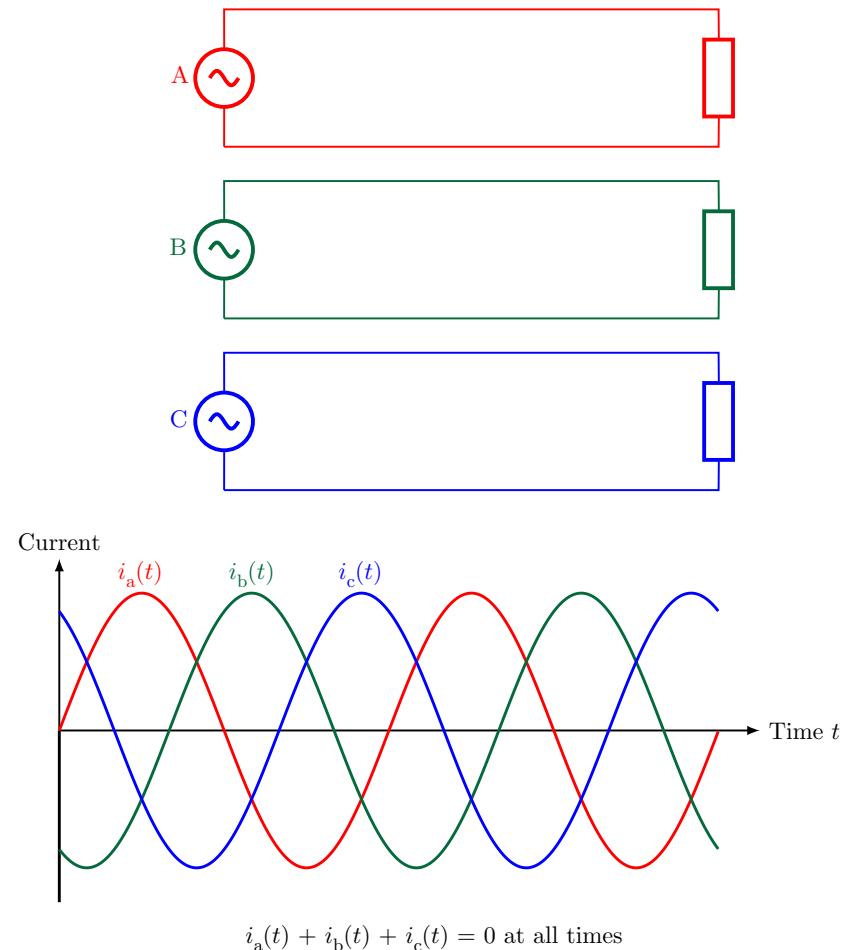


Figure 4.1 Three balanced single-phase a.c. currents.

choose to shift each by one-third of a cycle, or 120° . We now take the bold step of combining the return conductors of the three separate circuits, as shown in Figure 4.1.

Crucially, we assume that the magnitudes of a.c. voltage and current flow are equal in all three, except that their timing is shifted by 120° . Now, what happens in the combined return wire?

First, wouldn't we cause some awful short circuit by physically touching these three wires together? That is a good intuition, but in fact we can get away with making such a connection, provided that there is no other connection anywhere else in any of the three circuits that would define their potential or voltage relative to each other. Notice that we have drawn the three independent circuits as floating, without reference to ground. The voltage source in each circuit mandates a relative voltage difference between points on each individual circuit, but it says nothing about how this relates to any other circuit, or to the ground. Therefore, we are free to choose exactly *one* reference point on each independent circuit. We choose the point at one end of each of the three loads and combine the three circuits into one, as shown in Figure 4.2. This connection point is called the *neutral*.

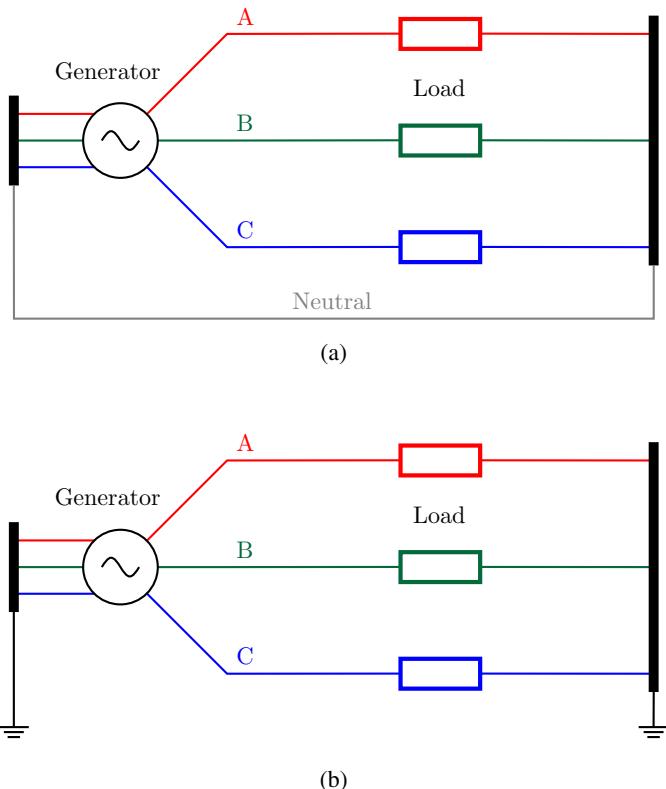


Figure 4.2 Three phases with and without neutral return; (a) three phases with common return; (b) three phases with neutral removed.

Physically gathering up the three separate wires is equivalent to bringing them to the same potential, but also adding the currents through them. In phasor notation, we can write²

$$I_A + I_B + I_C = I_N$$

where I_N is the current in the neutral conductor.

Consider now the sum of the three sinusoidal, phase-shifted currents at any particular instant. Trigonometric identities can prove that

$$\cos(\omega t) + \cos(\omega t + 120^\circ) + \cos(\omega t + 240^\circ) = 0$$

for all t . This result is best shown graphically with phasors, where the sum of three current phasors will describe an equilateral triangle as long as the magnitudes are equal:

$$I_A \angle 0^\circ + I_B \angle 240^\circ + I_C \angle 120^\circ = 0 \quad \text{iff} \quad |I_A| = |I_B| = |I_C|$$

By choosing a phase shift of -120° for Phase B and $-240^\circ = +120^\circ$ for Phase C, we are adopting the standard *positive sequence* convention where, in counterclockwise rotation, Phase B lags behind A, and C lags behind B. All the math would work just fine if we switched the B and C labels, but we would risk massive confusion in Section 4.2 on sequence components.³

² As threatened in an earlier footnote, we relax the practice of using boldface notation for phasors.

³ The labels A B C or L1 L2 L3 are standard, although other naming conventions for three phases are sometimes encountered, such as R S T, red yellow blue, etc.

Many students find it more intuitive to consider the time-domain graph of three staggered sinusoids in Figure 4.1. In some places, it is obvious from inspection that the three curves add up to zero. For example, at the instant when current A is zero, current B is exactly the negative of current C. Or, when current A is at its maximum, B and C are both at one-half their negative maximum. Elsewhere, it is not as easy to see, but the math holds up: the sum of currents A, B, and C is always zero, anywhere along the graph. The phasor diagram is a perfect synopsis of these relationships for all times t . If we imagine the trio of phasors rotating, their projections onto the real axis will vary, but their sum remains zero as the positive and negative contributions always cancel.

Let us now return to voltage in the combined three-phase circuit. In reality, the three-phase voltages would come from three separate windings in a generator, which produce an alternating potential across them at different times (owing to their different orientation relative to a rotating magnetic field, see Sections 10.1 and 10.3). From symmetry inside the machine, we may assume that the three voltages have the same magnitude and are spaced exactly one-third of a cycle apart. The voltage phasors are analogous to the current phasors, and also add to zero:

$$V_A \angle 0^\circ + V_B \angle 120^\circ + V_C \angle 240^\circ = 0 \quad \text{iff} \quad |V_A| = |V_B| = |V_C| \quad (4.1)$$

A standard arrangement (known as a *wye connection*) is to connect these three generator windings together at a common neutral point, just as we described for the loads above. We now connect the neutral points at the generator and the loads by way of the combined return conductor. The neutral is also a natural point in the circuit to connect to ground. We should now expect zero voltage difference across this neutral conductor, and zero current to flow through it.

In that case, why not just get rid of the wire? Based on the above analysis, we can simply connect the neutral at the generator as well as the loads to ground, as shown in Figure 4.2b. No current will need to flow through the earth. Consequently, we are providing power to three circuits with only three conductors, instead of the six we would have expected.

It's important to remember that while the generator controls the voltages, the load impedances determine the currents. Our key assumption is that each of the three phase loads has the same impedance, including both resistance and reactance. This ensures not only the same current magnitude on each phase, but that the currents will retain the 120° spacing in time as set by the voltages, since the three currents will be equally leading or lagging. It also means the voltage drop across each load will be the same.

Realistically, load impedances on the three phases may not be perfectly equal at all times; this problem is addressed in Section 4.1.3. When loads are slightly imbalanced, the sum of their currents will not be exactly zero. Also, the neutral point by the loads will be floating at a potential slightly different from ground. This means that some current will flow from the neutral point where the three load circuits connect to the point where the circuit is physically grounded, with a piece of metal stuck in the dirt. This current will then travel through the earth to the ground at the source. Assuming the imbalance is small, the residual current through the earth impedance will have a negligible effect.

4.1.2 Number of Phases

Some readers may note that the clever scheme of halving the number of conductors in multiple circuits by combining their return is not unique to three phases. In fact, any number of phases, staggered evenly in time and with equal amplitudes, will share the same property that their (balanced) currents and voltages always add up to zero. For example, we could supply four circuits

with four wires for four phases, spaced 90° apart; five phases 72° apart; or even just two opposite phases, 180° apart. Why, then, choose three?

Two opposite phases would be physically problematic for two reasons. Inside the rotating generator, two phases would not allow for a steady torque over the course of the rotation. The constant mechanical torque on the generator rotor as the three staggered phases with their magnetic fields oscillate in their respective directions provides for a smooth and efficient transfer of energy, rather than a pulsation (see Chapter 10), much like a trained cyclist spins the pedals with continuous torque rather than just pushing down. Also, transmitting power with only two phases would mean a greater vulnerability to imbalances in the loading of the two circuits: if one circuit were loaded more heavily than the other, the effects on voltage and current would be more pronounced than they are with more phases to “absorb” the difference.

More than three phases, on the other hand, just becomes increasingly complicated and expensive. If all the power on the system were supplied in four phases, for instance, this would require that all transmission and major distribution lines consist of four separate conductors. In addition, each bank of transformers at every substation would have to comprise four instead of three transformers, and similarly for circuit breakers. Although for the same amount of total power the capacity of each of these four components could be proportionately smaller, their total cost would likely be greater due to economies of scale, and the planning effort for balancing loads would also increase.

It is worth mentioning an interesting advantage of using a larger number of phases, related to the construction of transmission towers. With more phases staggered evenly—say, six phases shifted by 60° each—the instantaneous potential difference between any two adjacent phases is smaller. This allows a closer physical spacing of conductors without risk of arcing between them, which means less width is required for the transmission tower. Such lines have in fact been built, but they’re rare, as the overall cost–benefit calculation still favors the smaller number of components at each end. Three turns out to be the most practical choice, and three-phase a.c. power systems have been a global standard since the early days of long-distance transmission.

Because the three phases are so interdependent, the set of three is described as a single *circuit*—in contrast to the nomenclature we used before, where we referred to each individual phase as a circuit. For example, a common style of transmission line with six total conductors carries two circuits. In drawings, each three-phase circuit can be represented by a single line. This perspective is generally applied in situations that assume balanced and trouble-free operation. It is the more complicated, less-than-ideal situations involving faults and unbalanced loads that call for a separate analysis for each phase.

4.1.3 Balancing Loads

In deriving the result that a circuit of three a.c. phases does not require a ground return conductor, we assumed that the impedance or total load connected to each phase is identical, making the amplitudes of the three currents equal and their phases still exactly 120° apart. But how is this accomplished in reality?

Certain loads such as large commercial motors are connected to all three phases and draw power equally from all of them. But electric service to most residential customers generally consists of only one phase.⁴ The challenge for the distribution planner is to allocate these customers among the phases as evenly as possible, so that the total connected load on each phase is about the same.

Where single-phase laterals branch out from the main distribution feeders, distribution engineers take care that the local areas served by these feeders comprise similar loads. If a three-phase feeder

⁴ This is true for the United States. In Europe, three-phase residential service is more common.

runs down a street, transformers that serve several houses each will tend to alternate among the phases, and when new customers are connected, the phase is chosen by considering the balance with other loads in the vicinity. However, the utility may have incomplete or inaccurate information on file about the phasing of their distribution system, and the equipment in the field may not be clearly labeled. For example, if service was restored in a hurry after a major storm that knocked down multiple distribution poles in an area, it would not be surprising if a customer formerly on Phase A were reconnected to Phase B, with no formal record or mapping of the change.

Even with perfect system maps, the physical equipment limits the precision with which loads can be balanced, because loads come in chunks and not in arbitrarily small increments. The balance will also fluctuate in real-time as customers connected to different phases turn their appliances on and off. In practice, then, balancing loads is a rather approximate procedure, unlike the idealized, perfectly balanced phases encountered in engineering texts.

Local imbalances in current or power delivered by each phase can be in the tens of percent. However, this variation will be greatest near the customers, and will diminish at higher levels of aggregation (i.e., the transmission system) by virtue of statistics, since the local imbalances are more likely to cancel each other out as many loads are added together.

When the loads on all phases are not equal, the currents no longer add to zero. This means that some return flow of current is needed. If only three conductors are provided for three phases, the return current must travel through the ground. Along distribution feeders where more pronounced imbalances are to be expected, a fourth neutral wire is sometimes provided for just this purpose.

Imbalance in the loads or currents will also have an effect on voltages. This is because of the property called *source impedance* of the system. As current increases, so does the voltage drop across any transmission or distribution line and any transformer, as these elements all have finite, nonzero impedances. Thus, even if the generator is an ideal voltage source, the voltage seen by the load will have some dependence on current. (This effect is familiar from seeing our lights dim momentarily when a big motor kicks in with a high starting current, increasing the voltage drop across the wiring in our house.) Luckily, the source impedance is usually low enough that the voltage will only be impacted by a few percent of the change in current. If current imbalance is itself only a few percent, we can expect three-phase voltages to be quite well balanced.

4.1.4 Delta and Wye Connections

An individual single-phase load connects to two conductors to make a closed circuit. When power is delivered in three phases, the designer has two distinct choices: a load can be connected either between one phase and a (typically grounded) neutral point, or between one phase and another phase. In each case, the three phases will supply three separate loads.

The first arrangement in Figure 4.3, in which three loads are connected between one pair of phases each, is called a delta connection because the schematic resembled the Greek letter Δ . The second arrangement, where three loads are connected between each phase and neutral, is called a wye connection as the schematic resembles the letter Y.

In the wye connection, the voltages are called phase-to-neutral or line-to-neutral, often labeled L-N.⁵ Here it is easy to see that voltages and currents for each load will be spaced 120° apart. The same is true for the delta connection, but it is less obvious. The key is to remember that the voltage “across” a load is in fact the difference between the voltage on one side and the other. Thus,

⁵ The terms “phase” and “line” are interchangeable in this context, as they both refer to the A, B, C conductors. The term “neutral” is distinct from ground.

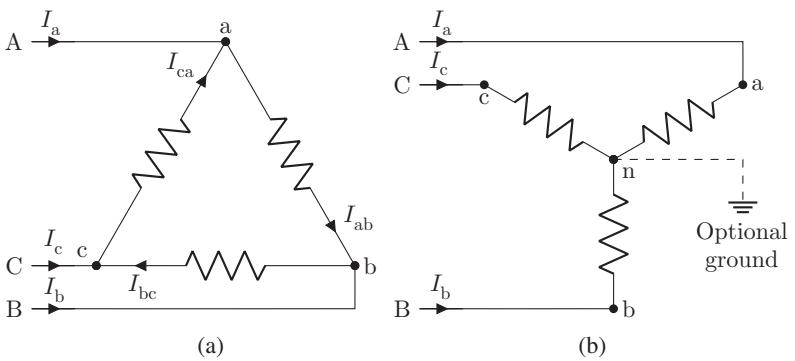


Figure 4.3 (a) Delta and (b) wye connections.

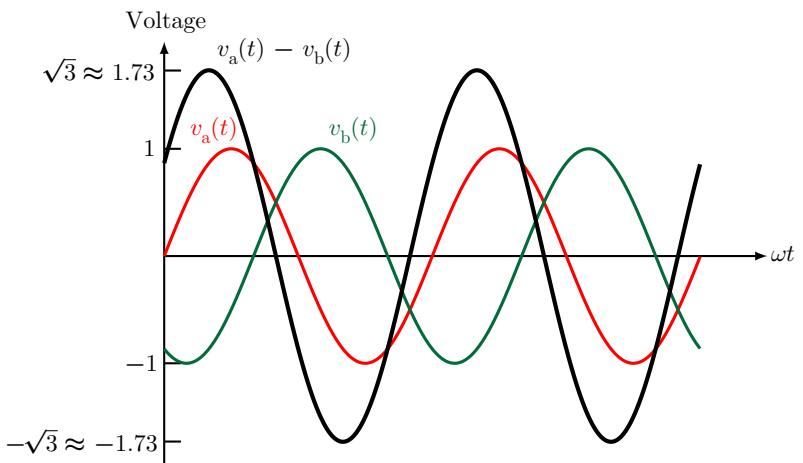


Figure 4.4 Phase-to-phase (line-to-line) voltage, seen in the time domain as the difference between a pair of phase-to-neutral (line-to-neutral) voltages.

mathematically, we can plot a curve that shows the difference between each pair of phase voltages at any moment in time, which gives the phase-to-phase voltage as a function of time (as shown in Figure 4.4).

Each of these three possible curves (a–b, b–c, or c–a) is still sinusoidal, but is shifted in time (phase) from the original phases a, b and c by 30° . It also has a different amplitude. As can be derived from trigonometric identities, the magnitude of the line-to-line voltage is (perhaps counterintuitively) greater than the line-to-neutral voltage by a factor of $\sqrt{3}$ (about 1.73).

The specific relationships are as follows:

$$V_{ab} = V_{an} - V_{bn} = \sqrt{3} V_{an} \angle 30^\circ$$

$$V_{bc} = V_{bn} - V_{cn} = \sqrt{3} V_{bn} \angle 30^\circ$$

$$V_{ca} = V_{cn} - V_{an} = \sqrt{3} V_{cn} \angle 30^\circ$$

These relationships are more elegantly displayed in a phasor diagram. In Figure 4.5, the line-to-line voltage phasors V_{ab} , V_{bc} , and V_{ca} are seen as the difference between the respective pairs of line-to-neutral voltages V_{an} , V_{bn} , and V_{cn} . Because phasors, like any vectors, retain their meaning

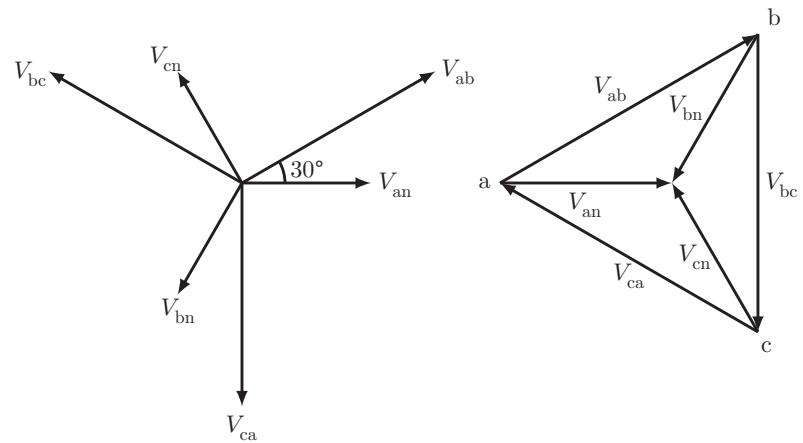


Figure 4.5 Line-to-line and line-to-neutral voltages represented in the phasor domain. The two different graphic arrangements are functionally identical: although the right diagram appears to “connect” voltages of the same value to support intuition, a phasor diagram is not a circuit diagram.

when moved across their plane (translation only, not rotation), their respective placement on the diagram is arbitrary. Depending on context, it may be convenient to depict the phasors either as having the same origin (left-side diagram), or in a more physical mapping (right-side diagram). Either style of phasor diagram shows the phase shift and magnitude difference arising from the vector subtraction.

On the right side of Figure 4.5, the neutral in the center is not necessarily at 0 V relative to ground, but it is a common point of physical connection for the L-N voltages. Likewise, the L-L voltages describe a physical connection between a pair of end points of the L-N phasors. If the three-phase voltages are imbalanced, the outer triangle described by the L-L voltages will be distorted and no longer exactly equilateral, but it is by definition a closed triangle.

An analogous diagram can be drawn for current phasors. Let us first be careful about defining the reference directions, shown in Figure 4.6. The term “line current” refers to the current in any one phase. In the wye connection, the line current I_a , I_b , or I_c flows through the load to neutral, and all three line currents are defined as positive going toward neutral. In the delta connection, the positive reference direction follows a cyclical pattern. The relationships between the line currents and the current in each branch of the delta can be obtained from by writing down Kirchhoff’s current law

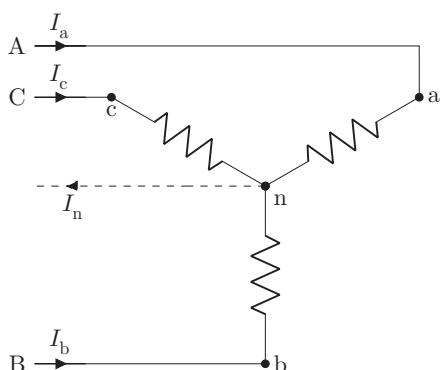
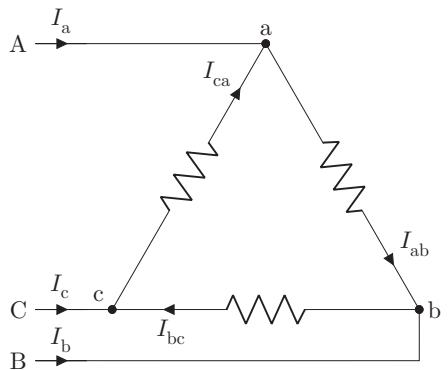
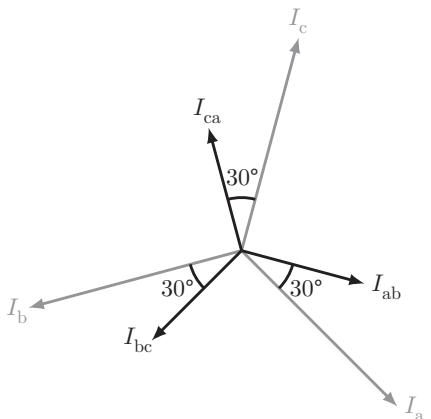


Figure 4.6 Currents in a wye connection.

Figure 4.7 Currents in a delta connection.**Figure 4.8** Current phasors showing the relationships between line currents (I_a , I_b , and I_c) and the currents through a delta-connected load (I_{ab} , I_{bc} , and I_{ca}).

(KCL) for each node, referring to the right side of Figure 4.7:

$$I_a = I_{ab} - I_{ca} = \sqrt{3} I_{ab} \angle -30^\circ$$

$$I_b = I_{bc} - I_{ab} = \sqrt{3} I_{bc} \angle -30^\circ$$

$$I_c = I_{ca} - I_{bc} = \sqrt{3} I_{bc} \angle -30^\circ$$

Once again we have the factor of $\sqrt{3}$ and a 30° shift from the subtraction. These relationships are best seen in the phasor diagram as shown in Figure 4.8. Note that the current in each branch of the delta is *less* than the line current.

4.1.5 Practical Aspects

The choice of delta or wye connection offers one way to select different voltage levels for loads from the same three-phase power supply. This is, in fact, how some utility customers, usually in the industrial or commercial sector, obtain both 120 V and so-called “220 V” from two or three phases and a neutral conductor. Actually, the higher (line-to-line) voltage in this arrangement is $120 \text{ V} \cdot \sqrt{3} = 208 \text{ V}$.⁶ The lower-voltage loads, such as lights and regular plug loads, would be connected between either phase and neutral, preferably balanced evenly among the three phases

⁶ An alternative approach for supplying dual voltages when only a single phase is present, called split-phase supply, involves transformer taps and is discussed in Section 6.2.

as this will reduce current flow in the neutral. Heavier single-phase loads could be connected between any pair of phases, and larger motors to all three phases, at 208 V. In these situations, care must be taken to match the voltage to the load rating, as some but not all appliances labeled “220 V” or “240 V” tolerate a 208-V supply. Another common industrial and commercial three-phase supply standard is 277/480 V (note the factor of $\sqrt{3}$).

In general, sources and loads in power systems are connected through transformers. Because the wire coils in the transformer always involve pairs of conductors that constitute each single-phase circuit, every three-phase transformer has to have either a delta or a wye configuration; both are common.

Aside from the difference in voltage level, the choice of delta or wye connection has some ramifications for reliability, in case there should be a short circuit. The delta configuration in general can be ungrounded or floating, meaning that no point on the circuit is connected to ground or to any point that has a specific defined potential. The (time-varying) potential differences between and among the points on the circuit are all that matters. The entire delta system can thus float at an arbitrary potential relative to ground. Different practices and standards apply in different countries for using floating delta connections, or referencing the delta to ground in a particular way.⁷

A floating delta circuit can continue to operate on an emergency basis, if any part of the circuit accidentally comes in contact with ground, which would otherwise cause a short-circuit and power outage. It is crucial, though, to detect and fix the problem before a second fault occurs elsewhere on the circuit, because this could now cause a dangerous situation with a high fault current.

Because of this property, the delta configuration is typically used where reliability is most important, such as auxiliary equipment in power plants, or on smaller transformers. The wye configuration, by contrast, is normally grounded at the center or neutral point. Here, a single ground anywhere else in the system will immediately cause a fault, and ground relay protection is always used to open the circuit breakers in such an event to protect the lines and equipment. In this case, the risk of damage to equipment overrides the short-term reliability concern. The wye connection is typically used on generators, main transformer banks, and transmission lines. As discussed in Section 8.3, transformers may be wired in a delta configuration on one side and wye on the other.

4.1.6 Three-phase Complex Power

To calculate the power transmitted on a three-phase transmission or distribution line, we essentially take the product of voltage and current for each individual phase and multiply by three. The only trick is taking into account a factor of $\sqrt{3}$ in just the right places.

Suppose we begin by assuming (incorrectly) that power is the product of line-to-line voltage and line current. We must reduce this estimate by a factor of $\sqrt{3}$ in the case of a wye connection, because the effective voltage seen by the load is the line-to-neutral, not the line-to-line voltage. In the delta connection, the power is reduced by a factor of $\sqrt{3}$ because the current seen by the load is not the line current.

For the wye connection, apparent power delivered by the *entire three-phase circuit* is given by

$$S_Y = 3 V_{LN} I = 3 \left(\frac{V_{LL}}{\sqrt{3}} \right) I = \sqrt{3} V_{LL} I \quad (4.2)$$

where $V_{LN} = |V_{an}| = |V_{bn}| = |V_{cn}|$ is the rms line-to-neutral voltage, $V_{LL} = |V_{ab}| = |V_{bc}| = |V_{ca}|$ is the rms line-to-line voltage, and I is the rms line current in any one phase conductor. Note that we

⁷ An excellent practical discussion of different labeling and grounding standards can be found in Alex McEachern, “Designing Equipment for World-Wide Power,” *Conference Record of the 2002 IEEE Industry Applications Conference*. 37th IAS Annual Meeting (Pittsburgh, PA: IEEE, 2002).

are dealing here with magnitudes only, since we are describing an entire three-phase circuit and not making explicit reference to the timing of any one phase.

For the delta case, using $I_{\Delta} = |I_{ab}| = |I_{bc}| = |I_{ca}|$ for the branch current between each pair of phases, we arrive at exactly the same result:

$$S_{\Delta} = 3 V_{LL} I_{\Delta} = 3 V_{LL} \left(\frac{I}{\sqrt{3}} \right) = \sqrt{3} V_{LL} I \quad (4.3)$$

To obtain three-phase real and reactive power, we simply apply $\cos \theta$ (the power factor) and $\sin \theta$, respectively, just as in the single-phase case. Since this framework assumes balanced loads, the power factor must be the same for each phase.

4.1.7 Three-phase Impedance

It is important to recognize that we can determine the power delivered from the line voltage and line current, without having to know how the load is connected (delta or wye). This stands to reason because we could not gain or lose energy simply by reconfiguring the connection, without changing voltage and current in the transmission line.

However, if we took a wye load and reconnected it in a delta configuration, or *vice versa*, it would change the physics of the situation. We know this because we correctly think of most loads as having a particular, fixed impedance. The statement that three-phase power is independent of delta or wye connection tacitly implies that the load would be designed with a different impedance for the two cases.

Specifically, we get the same total power draw for a three-phase wye impedance Z_Y connected line-to-neutral, or a delta impedance Z_{Δ} connected line-to-line, if $Z_{\Delta} = 3 Z_Y$. We can demonstrate this result by writing Ohm's law for each case:

$$V_{LN} = I Z_Y$$

$$V_{LL} = I_{\Delta} Z_{\Delta}$$

Then if $V_{LL} = \sqrt{3} V_{LN}$ and $I = \sqrt{3} I_{\Delta}$, it must be true that $Z_{\Delta} = 3 Z_Y$.

This result is intuitive if we consider a load with Z_Y (designed for the line-to-neutral voltage) being erroneously connected to the line-to-line voltage in a delta configuration. By Ohm's law, its current would proportionally increase by $\sqrt{3}$ along with the $\sqrt{3}$ times greater voltage, to draw three times as much power (and probably do some damage). Conversely, a load with Z_{Δ} connected to the line-to-neutral voltage would draw only 1/3 the intended power. Voltage ratings on electrical devices make an implicit statement about whether the internal impedance is sufficiently high to be safely connected to that voltage.

4.2 Symmetrical Components

Symmetrical components is the summary term for *positive-, negative-, and zero-sequence* components. The term "component" refers to sets of voltage or current phasors within a three-phase system that can be mathematically decomposed in different ways.

In describing balanced three-phase systems, we have assumed purely positive-sequence voltages and currents. We briefly mentioned the problem that if the loads on each phase are not in fact balanced, the currents and even the voltages will not match the idealized description. Symmetrical components provide a rigorous way to account for phase imbalance, for use in specific situations

where the imbalance is significant and consequential. Most often, this means an abnormal operating condition or fault in the system. For example, power engineers must have a way to calculate fault currents for the case where only one or two phases are faulted. Symmetrical components can also be used to perform detailed calculations of voltages and currents on distribution systems where single-phase customers or distributed generation are not evenly connected among the three phases. While these methods are beyond the scope of this text, we introduce symmetrical components here to help the reader make sense of the term “positive sequence” that is often encountered in practice.

The key to the method is that instead of accounting for each phase individually, we can describe the entire unbalanced situation as a combination of three balanced situations. Specifically, the set of three ABC phasors can be represented as a linear combination or vector sum of three phasor trios, each of which is internally balanced or symmetrical.⁸ In linear algebra, we would say that the symmetrical components offer a different set of basis vectors in which to express our objective.

While the conversion to and from symmetrical components may seem tedious at first, working in terms of these components greatly simplifies the arithmetic for unbalanced three-phase systems. The reason is that we cannot consider any one phase individually without the other two, as they are coupled both electrically and magnetically (see Section 9.1). To account for all the mutual dependencies, we would be stuck with not three but nine voltage–current relationships. The method of symmetrical components cleverly summarizes all the necessary information and allows us to write only three equations to characterize the entire system, even when the three phases are not behaving the same way.

Symmetrical components are best introduced by visual example. Figure 4.10 illustrates how an unbalanced set of phasors (A , B , C) is produced by vector summation from three sets of symmetrical components, from the three sets of symmetrical components as phasors in the complex plane. They could be either currents or voltages. Each of these phasors on the right-hand side of the diagram is the sum of three components: namely, the a , b , and c components⁹ of three symmetrical trios of vectors that are shown separately in Figure 4.9.

The first trio is called *positive sequence* and looks exactly like a balanced set of A , B , C phasors: the magnitudes are equal, the angle spacing is 120° , and the rotation is counterclockwise so that component b lags a , which lags c .

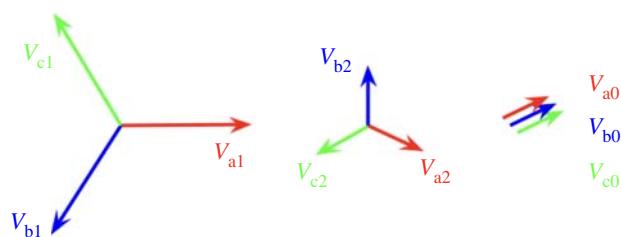


Figure 4.9 Positive-, negative-, and zero-sequence components. Each trio includes three phasors of equal magnitude, imagined as rotating counterclockwise.

⁸ This is also known as Fortescue’s method, after a classic 1918 paper by Charles Legeyt Fortescue that proved any set of N unbalanced phasors could be expressed as the sum of N symmetrical sets of balanced phasors, for values of N that are prime.

⁹ There is no standard convention for upper- or lowercase phase labels; the distinction is made here for clarity.

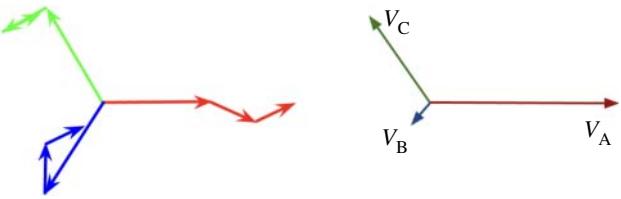


Figure 4.10 Vector addition of the symmetrical components in Figure 4.10, corresponding to Eq. 4.6.

The second trio is *negative sequence*. Again, the magnitudes are equal and the spacing is 120° , but the sequence is now reversed such that, under counterclockwise rotation, component b leads a, which leads c.¹⁰

The third trio, called *zero sequence*, has all three of its a, b, c components coincide exactly. Now, if we add together the a-components of the positive-, negative-, and zero-sequence trios, we obtain the original A phasor. Likewise, summing up the three symmetrical b-components yields the original B phasor, and the same for c.¹¹

For a balanced three-phase system, the positive-sequence components in and of themselves completely describe the situation, and the negative- and zero-sequence components vanish. The remarkable fact is that any kind of phase imbalance—including magnitudes as well as angle shifts—can be characterized by suitable choice of magnitude and relative angle for the negative- and zero-sequence trios.

For example, in the imbalanced case illustrated in Figure 4.10, phasor A has a greater magnitude than B or C. Adding some negative-sequence contribution to the balanced case, with the a-component angles of the negative- and positive-sequence trios roughly aligned, increases the A magnitude of the sum. At the same time, there is some mutual cancellation of b and c components that reduces the resulting Phase B and C magnitudes and also shifts their relative angles to something other than 120° spacing. Finally, some zero-sequence contribution could be added to the recipe to further alter the relationship among the three phases. Note that the situation depicted here is (hopefully!) a gross exaggeration of phase imbalance seen in practice.

It is not intuitively obvious that a recipe should exist for every possible case of unbalanced phases. We can see, however, that the amount of information contained in either format is the same. Consider that any set of three phasors—not balanced, but of the identical frequency—is specified by six pieces of information: three magnitudes, and three phase angles. The same is true for three symmetrical component trios. Since each trio consists of three vectors that by definition have the same magnitude and a known spacing relative to each other, each trio is completely characterized by any one of its constituents. By convention, we use the a-component to indicate the relative angle. Thus, each trio is specified by two pieces of information, also making six pieces total.¹²

¹⁰ The same behavior could be construed by leaving the components in their previous positions and letting them rotate clockwise instead, but this would be mathematically improper.

¹¹ An animated visualization of symmetrical component addition in both the time and phasor domain is offered by Alex McEachern's Power Quality Teaching Toy, <https://mcelabs.com/powerqualityteachingtoy>.

¹² In either representation, one of the three angles amounts to a choice of coordinate system (for example, the angle of the Phase A voltage or the a-component of the positive-sequence trio at time $t = 0$). So there are really only five variables that characterize the physical situation. The sixth piece of information only serves as a reference, which might allow comparison to a different set of phasors in the same coordinate system. In other words, the entire Figure 4.11 can be rotated without loss of meaning.

4.2.1 Converting Symmetrical Components

For doing arithmetic with symmetrical components, we use the conventional subscripts 0, 1, 2 for zero, positive, and negative sequences, respectively.¹³ The ensemble is written¹⁴ as

$$\mathbf{V}_{012} = \begin{bmatrix} V_0 \\ V_1 \\ V_2 \end{bmatrix} \quad (4.4)$$

We also introduce a shorthand notation, the operator α , for rotating a vector by 120° in the complex plane. Specifically, we define:

$$\alpha \equiv e^{\frac{2}{3}j\pi} = 1\angle 120^\circ \quad (4.5)$$

Note that $\alpha^2 = -\alpha$ (applying the rotation operator twice is the same as applying it backwards) and $\alpha^3 = 1$ (applying it three times gets you back to where you started).

We use voltage for illustration purposes, but all the analogous relationships hold for current.¹⁵ We can now write the symmetrical components as

$$\begin{aligned} V_0 &\equiv V_{a,0} = V_{b,0} = V_{c,0} \\ V_1 &\equiv V_{a,1} = \alpha V_{b,1} = \alpha^2 V_{c,1} \\ V_2 &\equiv V_{a,2} = \alpha^2 V_{b,2} = \alpha V_{c,2} \end{aligned} \quad (4.6)$$

and the original phasors as the sum of their respective symmetrical components:

$$\begin{aligned} V_A &= V_{a,0} + V_{a,1} + V_{a,2} \\ V_B &= V_{b,0} + V_{b,1} + V_{b,2} \\ V_C &= V_{c,0} + V_{c,1} + V_{c,2} \end{aligned} \quad (4.7)$$

The entire system can be expressed concisely in matrix notation:

$$\mathbf{V}_{ABC} = \begin{bmatrix} V_A \\ V_B \\ V_C \end{bmatrix} = \begin{bmatrix} V_0 \\ V_0 \\ V_0 \end{bmatrix} + \begin{bmatrix} V_1 \\ \alpha^2 V_1 \\ \alpha V_1 \end{bmatrix} + \begin{bmatrix} V_2 \\ \alpha V_2 \\ \alpha^2 V_2 \end{bmatrix} \quad (4.8)$$

which condenses to

$$\mathbf{V}_{ABC} = \mathbf{A} \mathbf{V}_{012} \quad (4.9)$$

where we define

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha^2 & \alpha \\ 1 & \alpha & \alpha^2 \end{bmatrix} \quad (4.10)$$

¹³ The subscripts 0, +, – would be more obvious but are less commonly used, perhaps because the minus sign can be hard to read.

¹⁴ In this section, we use boldface notation to distinguish matrices from individual phasors.

¹⁵ The letter V just has more visual appeal. In practice, currents would be more likely to suffer significant imbalance than voltages.

Conversely, we can obtain the symmetrical from the phase components by inverting the relationship:

$$\mathbf{V}_{012} = \mathbf{A}^{-1} \mathbf{V}_{ABC} \quad (4.11)$$

where

$$\mathbf{A}^{-1} = \frac{1}{3} \begin{bmatrix} 1 & 1 & 1 \\ 1 & \alpha & \alpha^2 \\ 1 & \alpha^2 & \alpha \end{bmatrix} \quad (4.12)$$

The scaling factor of 1/3 in the inverse matrix makes sense because we defined the ABC phasors as the sum of three symmetrical components, rather than the other way around. Therefore, if we now want to express symmetrical components as the sum of ABC phasors, we should only take a third of the magnitude of each. In a sense, we are averaging the ABC components. Many students will find that the matrix notation does less for building intuition than the graphic visualization, but it supports the mechanics of converting between formats.

Given a set of voltage phasors \mathbf{V}_{ABC} , writing Eq. (4.11) out in longhand gives the positive-, negative-, and zero-sequence components:

$$V_0 = \frac{1}{3}(V_A + V_B + V_C)$$

$$V_1 = \frac{1}{3}(V_A + \alpha V_B + \alpha^2 V_C)$$

$$V_2 = \frac{1}{3}(V_A + \alpha^2 V_B + \alpha V_C)$$

The zero-sequence component is the average of the ABC phasors. If the original set were balanced, they would cancel in the vector sum. For the positive sequence, we are effectively projecting the B and C phasors over toward the neighborhood of the A phasor. If their spacing had been exactly 120°, they would now overlap and we would only average their magnitudes. For the negative sequence, we are projecting the B and C phasors to roughly switch positions, and then taking the average; if the original set were balanced, this would yield zero.

Example

A slightly imbalanced set of loads produces the following set of current phasors:

$$I_A = 10\angle 0^\circ$$

$$I_B = 9\angle -125^\circ$$

$$I_C = 11\angle 119^\circ$$

This type of situation commonly arises when three loads are presented with a balanced three-phase voltage but have somewhat different impedances, including magnitude as well as power factor that shifts the angle of the respective phase currents.

The positive sequence is defined by its a component:

$$I_1 = I_{a,1} = \frac{1}{3}(10\angle 0^\circ + 9\angle -5^\circ + 11\angle -1^\circ) = 9.99\angle -1.9^\circ$$

This should resemble I_A quite closely, because the phases were not terribly far from being equal in magnitude and rotated by 120°. The above answer for $I_{a,1}$ also implies that

$$I_{b,1} = 9.99\angle -121.9^\circ \quad \text{and} \quad I_{c,1} = 9.99\angle 118.1^\circ$$

Since the phase imbalance is fairly minor, we would expect the negative- and zero-sequence components to be small. We obtain the following negative-sequence components:

$$I_2 = I_{a,2} = \frac{1}{3}(10\angle 0^\circ + 9\angle 115^\circ + 11\angle -121^\circ) = 0.177 - j0.423 = 0.459\angle -67.3^\circ$$

It is useful to write out the real and imaginary parts to confirm which quadrant the resulting angle should lie in, since the inverse tangent has two possible solutions. The $I_{a,2}$ component completely defines the trio, and we can write

$$I_{b,2} = 0.459\angle 52.7^\circ \quad \text{and} \quad I_{c,2} = 0.459\angle 172.7^\circ$$

noting the reversal of order in the counterclockwise sense.

Finally, the zero-sequence components are

$$I_0 = I_{a,0} = I_{b,0} = I_{c,0} = -0.165 + j0.750 = 0.768\angle 102.4^\circ$$

Here we are careful to choose the solution in the second quadrant (instead of -77.6° , which our calculators might offer as the default answer). This is consistent with the intuition that the zero sequence should point roughly in the direction of the largest of the original ABC phasors, and thus somewhat closer to I_C than the others. We could check our work by confirming that $I_{a,1} + I_{a,2} + I_{a,0} = I_A$, and similarly for the other phases.

4.2.2 Ohm's Law with Symmetrical Components

The benefit of symmetrical components is that they allow us to describe an imbalanced three-phase situation in terms of three separate, superimposed balanced situations, instead of having to explicitly account for the dependence of voltage and current quantities across three phases.

Somewhat analogous to what we did for voltage and current, we can define a trio of symmetrical impedances that captures the impedances for all three phases, decomposed into a positive-, negative-, and zero-sequence component, where

$$\begin{aligned} V_0 &= I_0 Z_0 \\ V_1 &= I_1 Z_1 \\ V_2 &= I_2 Z_2 \end{aligned} \tag{4.13}$$

The intuitive interpretation is that because Ohm's law applies at every instant, currents of any given sequence (positive, negative or zero) can only be associated with voltage drops of the same sequence. The properties of the three-phase conductors and their magnetic interactions are distilled into three impedance terms, one for each current-voltage relationship.¹⁶

For transmission lines and three-phase transformers, we expect the positive- and negative-sequence impedances to be the same, that is, $Z_1 = Z_2$. The zero-sequence impedance is physically different because zero-sequence currents flow in the same direction at the same instant. This impacts not only the mutual inductances between phases, but also the return path for the current. The return path must now involve the neutral or ground, which tends to be spaced farther away from the phase conductors. Therefore, we generally expect the zero-sequence impedance to be greater than the positive- and negative-sequence impedances. In rotating machines, positive- and negative-sequence impedances are also distinct from each other, because the direction of physical rotation matters.

¹⁶ Note that this still assumes some symmetry between each of the three-phase conductors and their mutual spacing. The completely general model of coupling among unbalanced phases is rarely used in practice, since the accuracy of information from the field (e.g., exactly how many inches between conductors) is usually too poor to justify the computational effort.

Analysis with symmetrical components is mostly done for protection engineering, where highly unbalanced fault currents are expected. Outside this context, it is common to hear only about the positive-sequence impedance, with the other components neglected. For example, when working with one-line diagrams that represent three-phase transmission or distribution lines, we assume a balanced three-phase system and use only the positive-sequence terms.

4.3 Direct and Quadrature Components

The *DQZ* or *dqz* transformation converts a set of rotating phasors into a set of stationary vectors in a rotating reference frame. This is especially useful for the analysis of voltages and currents within three-phase machines. It offers a very different kind of simplification than symmetrical components: instead of eliminating imbalance, it eliminates common movement from the analysis. This allows representing alternating currents as direct currents. The term *dqz* stands for three axes in a rotating coordinate system: *direct*, *quadrature*, and *zero*.

The direct and quadrature axes refer to the physical position of a generator rotor (Section 10.1), or its virtual equivalent in an inverter (Section 14.4). Specifically, the direct or *d*-axis is defined as pointing directly in the direction of the rotor, assumed to be spinning at the nominal frequency (50 or 60 Hz).¹⁷ Quadrature is a somewhat archaic term associated with finding an area or square; in power engineering, it has traditionally been used to refer to quantities at right angles from each other. For example, a current lagging a voltage by 90° is said to be “in quadrature.” This is a useful term because when decomposing any current into a direct and a quadrature component, we know that the quadrature part involves no real power transfer. The *q*-axis is thus defined as pointing at 90° from the *d*-axis. By convention, the positive *q*-axis leads the *d*-axis in the sense of rotation. The *z*-axis is orthogonal to the plane of rotation.

Mathematically, the *dqz* transformation combines a geometric projection of vectors onto another set of basis vectors (the *Clarke transform*) with a rotation of that coordinate system (the *Park transform*). It answers the question, how much of a given phasor is consistently aligned with the orientation of a generator rotor? This is particularly useful in the context of control problems, where the focus is on deviations from a steadily moving reference. For example, one might want to quantify if a current or voltage is pulling ahead or falling behind the reference rotation, which would be associated with forces on the machine and changes in the amount of instantaneous power. The *dqz* transformation allows expressing relevant quantities in terms of a *linear, time-invariant system* that is amenable to certain techniques in control theory.

The related sections in this book (Sections 10.3, 10.4, 10.6, and 13.4) are written without reference to direct and quadrature components because for the majority of readers, doing so would likely increase rather than decrease the cognitive load. Readers who actually need to work in *dqz* terms should be well prepared, based on this qualitative introduction, to interpret matrices found in standard references.

Problems and Questions

- 4.1** A common transmission line voltage is labeled as “230 kV,” referring to the root-mean-square (rms) line-to-line voltage.
 (a) What is the rms line-to-neutral voltage?

¹⁷ As detailed in Sections 10.1 and 10.3, the physical position of the rotor and its magnetic field relates to the voltage produced by the machine at any given instant.

- (b) What is the maximum instantaneous voltage difference between any two phases?
- (c) What is the maximum instantaneous voltage difference between any one phase and the transmission tower? Why would you care about this particular quantity when designing the transmission structure?
- 4.2** The German word for three-phase a.c. is *Drehstrom*, which literally translates to “turning current.” Explain in your own words how this terminology makes sense.
- 4.3** Three identical impedances $Z = 20\angle 30^\circ$ are connected in a wye-configuration, where the line-to-neutral voltages are $V_A = 100\angle 0^\circ$ V, $V_B = 100\angle -120^\circ$ V, and $V_C = 100\angle 120^\circ$ V.
- Find the currents I_A , I_B , and I_C .
 - Find the total three-phase power $|S_Y|$.
 - Suppose the identical impedances Z are now connected instead in a delta (line-to-line) configuration across these phases, which still carry the same voltage as before. Find the current I_{AB} , I_{BC} , and I_{CA} in each load.
 - Find the total three-phase power $|S_\Delta|$.
 - What would the impedance Z_Δ for the delta connection have to be in order to yield the same power as in the wye connection before?
- 4.4** A three-phase load draws 240 kW at a power factor of 0.707 lagging, from a 440-V line (labeled by convention in terms of the line-to-line voltage). For this problem, you don't have to know whether this load is connected in a delta or wye configuration. In parallel with the load is a three-phase capacitor bank that “supplies” 60 kVAR.
- Find the total current on any single phase of the line leading to the load–capacitor combination.
 - Find the power factor of the parallel combination.
 - Why don't you have to know whether it's delta or wye? What would be different in the two configurations, but doesn't affect your answer to (a) and (b)?
- 4.5** A three-phase, wye-connected motor draws 20 kVA at p.f. = 0.707 lagging from a 220-V source.
- Find the kVA rating of the capacitors to make the combined power factor 0.90 lagging.
 - Find the line current (on any one phase) before and after the capacitors are added.
- 4.6** Consider the delta configurations illustrated in Figure 4.11 (a) through (e). Suppose the line-to-line voltage in each case is 480 V. State the line-to-ground voltage you would expect to measure in each case for L1, L2, and L3, respectively, assuming the system is balanced.

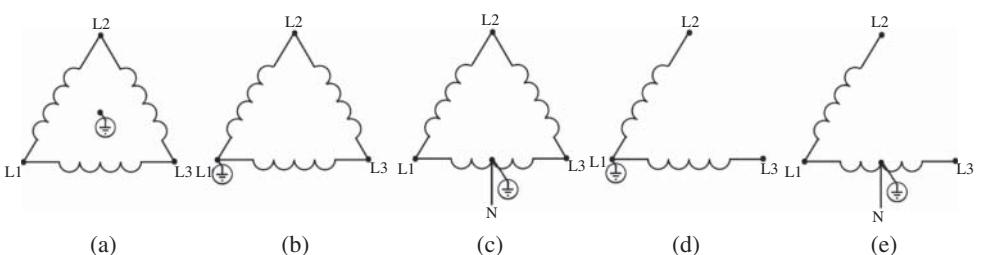


Figure 4.11 Various international practices for delta connections. Source: A. McEachern, ([2002] IEEE).

- 4.7** Consider an unbalanced set of three-phase currents corresponding to the symmetrical components $I_{a,1} = 1\angle 0^\circ$, $I_{a,2} = 1\angle 0^\circ$, and $I_0 = 0$. Find the resulting ABC phase currents, and sketch a phasor diagram.
- 4.8** Repeat the previous exercise for the set of currents corresponding to $I_{a,1} = 1\angle 0^\circ$, $I_{a,2} = 0$, and $I_0 = 0.5\angle 0^\circ$. Qualitatively, describe what happens if you change the relative angle of the zero-sequence component, I_0 .
- 4.9** Propose a set of positive-, negative-, and zero-sequence components for which you can write down the three-phase currents I_A , I_B , and I_C by inspection.
- 4.10** Consider an unbalanced set of three-phase currents described by the phasors $I_A = 10\angle 0^\circ$, $I_B = 9\angle -125^\circ$, and $I_C = 11\angle 119^\circ$. Find the positive-, negative- and zero-sequence components, and sketch a phasor diagram for both the symmetrical components and the original phasors.
- 4.11** An inventor asks your opinion about his new proposed five-phase a.c. system for an island nation.
(a) Is it plausible to build such a system? What issues or concerns do you raise?
(b) In a balanced five-phase system with line-to-neutral voltage 10 kV, what would be the line-to-line voltage?

5

Power Quality

Power quality encompasses voltage, frequency, and waveform. From a theoretical point of view, good power quality can be taken to mean that the voltage supplied by the utility at the customer's service entrance is steady and within the prescribed range, that the a.c. frequency is steady and very close to its nominal value, and that the *waveform* or shape of the voltage curve versus time very much resembles the smooth sine wave from mathematics textbooks (a condition also described as the absence of *harmonic distortion*). In practice, however, it makes more sense to consider power quality as the *compatibility* between what comes out of an electric outlet and the load that is plugged into it.

The proliferation of electronic technologies over the past several decades has turned “electric load” into a wide spectrum of devices that vary tremendously in their response to changes in voltage, frequency, and waveform: what is perfect power quality for one appliance may be devastating to the next, or cause it to behave in unexpected ways.¹

Not surprisingly, how much power quality is needed and by whom—and how much money it is worth—is the subject of some controversy. Because so many of the factors and events that bear on power quality and performance are beyond the control of the utility or system operator, a realistic goal is to ensure a working compatibility between the system and the job it is expected to do. This includes a power system that delivers “clean” enough power for its customers’ loads, and conversely, loads that are robust enough to tolerate the “dirty” power that their grid can reasonably be expected to provide.

5.1 Voltage

The voltage received by a utility customer varies along with power flows in the transmission and especially the distribution system. Initially, generators inject their power at a fixed voltage magnitude, which would translate through several transformers into a fixed supply voltage for customers. But as consumption increases and line current increases accordingly, there is an increasing *voltage drop* along the power lines according to Ohm’s law (Section 1.2). This means that the difference between the voltage supplied at the generation end and that received by a

¹ It is worth noting that electronics, and especially their power supplies that interface with the grid, are often designed and tested by engineers in laboratories where power quality is better than at the consumer’s location out on a typical distribution circuit. Thus, absent a specific testing protocol for ensuring a new gadget’s tolerance for voltage, frequency, and waveform variations, the design engineers may have no idea they are building a fragile piece of equipment that will behave annoyingly, or worse, in real-life conditions. I am indebted to Alex McEachern for this and other helpful observations concerning power quality.

given load varies continuously with demand, both systemwide and local. The utility can take diverse steps to correct for this variance, primarily at the distribution level (see Section 7.4 on voltage control), but never perfectly. The traditional norm in the United States is to allow for a tolerance of $\pm 5\%$ for voltage magnitude, which translates into a range of 114–126 V for a nominal 120-V service.

Low voltage may result if a power system's resources are overtaxed by exceedingly high demand. This condition is sometimes informally called "brownout" because lights become dim at lower voltage. Aside from the nuisance, operation at low voltage can damage electric motors. Excessively high voltage, on the other hand, can also damage appliances simply by overloading their circuits. Incandescent light bulbs, for example, have a shorter life if exposed to higher voltages due to thermal stress of the filament.

Beyond the average operating voltage, power quality concerns voltage swells and sags, or sudden and temporary departures from normal voltage levels that result from disturbance events in the distribution system. Abrupt voltage changes can be caused by lightning strike or by large inductive loads connecting and disconnecting, but they are most often related to faults on nearby distribution circuits. Specifically, in the time interval between the appearance of a fault and its isolation by a fuse or circuit breaker, a fault current (which is much greater than the usual load current) results in a significantly greater voltage drop along the entire feeder, as well as other feeders connected to it in a radial distribution system (see Section 7.1 on distribution system design and Section 7.5.2 on protection). This time interval before the circuit protection actuates may be anywhere from a fraction of a second to several seconds, depending on the fault's impedance, and thus the magnitude of the fault current.

To describe temporary voltage increases, the term *swell* is generally used to denote a longer event, whereas an *impulse* would last on the order of microseconds. The word *spike*, though used colloquially and found in the common "spike protector," is frowned upon by experts because of its ambiguity, having been used historically to describe rather diverse electrical occurrences. It is easy to see how a voltage swell or impulse can actually damage loads: the proportionally increased current could overheat a small component inside an appliance, or an arc may form between components that are insufficiently insulated, and the equipment gets "fried." Power strips with "spike" or "surge protection" are advertised as a protection against this risk, although anecdotally, the incidence of consumer equipment damaged by voltage swells in industrialized countries appears to be quite low.

The more important job of a surge protector may be to mitigate temporary decreases in voltage—*sags* to Americans and *dips* to the British—that can cause electronic loads to shut off or otherwise behave strangely. One might expect voltage sags to be essentially a nuisance, noticeable as a brief dimming of lights and the occasional rebooting of a computer, but they surprisingly constitute the most common power quality problem by an order of magnitude. Owing to the large number of sensitive commercial and industrial loads, economic losses from voltage sags in the United States are estimated on the order of tens of billions of dollars per year.²

It is difficult to obtain empirical data on how often voltage sags and other power quality problems occur because electric utilities, which are in a unique position to measure and record them, are typically not keen on publishing this information. More than just a matter of image, there is an issue of accountability and potential liability. With lower-cost monitoring equipment and data analytics,

² An example that vividly illustrates this point is a carpet manufacturing plant, where the tension on each of the myriad threads is a function of the voltage across the control. We can easily imagine the mess, as well as the cost, when one thread tears or bunches up momentarily, and the entire roll of carpet has to be thrown out.

independent analyses could play an increasingly important role, especially in regions where power quality and reliability often pose considerable problems for the average customer.³

Of course, power quality and reliability depend on infrastructure investment, from adequate equipment sizing to the most expensive option of placing distribution systems underground (see Sections 7.1 and 13.2). Nevertheless, many factors that substantially affect power quality—from squirrels and winter storms to backhoes and drunk drivers—lie beyond a utility's control, putting them in the awkward position of being responsible for system performance, yet unable to make firm guarantees. The inherent problem of vulnerability to external factors also limits the feasibility of contractual agreements with customers willing to pay more for better power quality because an electric distribution system in the real world, although it can be made comparatively more robust, simply cannot be guaranteed to operate without any disturbances whatsoever.⁴ As a result, customers with important, sensitive loads generally have to look to power conditioning equipment on their own side of the utility service drop.

The key to satisfactory power quality is not perfection, but compatibility between the power provided and the power expected by the customer's loads. Expectations regarding voltage tolerance are usefully described by the ITIC curve in Figure 5.1, which outlines an envelope of voltage excursions and their respective duration that should be tolerable for equipment.⁵

In a nutshell, the curve shows that the larger the voltage excursion, the briefer its duration must be in order to avoid equipment damage or malfunction. Note that the time scale is logarithmic. For example, we would expect an electronic device to safely tolerate twice the nominal peak voltage from the wall outlet for at least 1 ms, or five times nominal for 0.01 cycle (0.167 ms at 60 Hz and 0.2 ms at 50 Hz), and indefinitely accommodate a voltage 10% above nominal. The expectation for tolerating low voltage is defined as riding through a voltage sag without shutting off (or rebooting a computer) if, for example, a voltage reduction by no more than 30% below nominal lasts for less than half a second. The guideline for manufacturers is that a low voltage of extended duration should not cause physical damage to the equipment, even if it causes a nuisance interruption. The electric utility is responsible for preventing overvoltages in the prohibited region, and typically would be assumed liable for damages in such an event.

5.1.1 Conservation Voltage Reduction

Power consumption by loads generally tends to increase with service voltage. Depending on the regulatory framework, utilities whose financial bottom line hinges on kilowatt-hour sales might find it advantageous to maintain a higher service voltage profile. Conversely, it is possible to reduce

³ For example, see Veronica Jacome, Noah Klugman, Catherine Wolfram, Belinda Grunfeld, Duncan Callaway, and Isha Ray, "Power Quality and Modern Energy for All," *Proceedings of the National Academy of Sciences of the United States of America* 116(33), 16308–16313, July 29, 2019. DOI: 10.1073/pnas.1903610116.

⁴ Alex McEachern offers this instructive analogy: Imagine an assistant who does a fantastic job in every respect, except for one very strange habit. Every so often, and completely unexpectedly, he comes up from behind and pokes you in the back with a knitting needle. Suppose this happens, on average, six times per year. Since the assistant is very competent otherwise and you wish to retain him, you ask whether you could perhaps raise his salary so that he would not poke you any more. The answer comes back that with a raise, he could reduce the number of knitting needle incidents from six to three per year. This deal does not seem very attractive! Similarly, a utility customer plagued by occasional voltage sags might be willing to pay more money for better power quality, but what they want is *zero* incidents, which the utility cannot realistically offer.

⁵ This curve was originally developed by the Computer Business Equipment Manufacturer's Association (CBEMA) in the 1970s as a guideline for designing power supplies and entered the mainstream vocabulary as a reference for equipment design and power quality. CBEMA has since become the Information Technology Industry Council (ITI), and the formerly smooth curve was revised to the piecewise linear envelope seen in Figure 5.1.

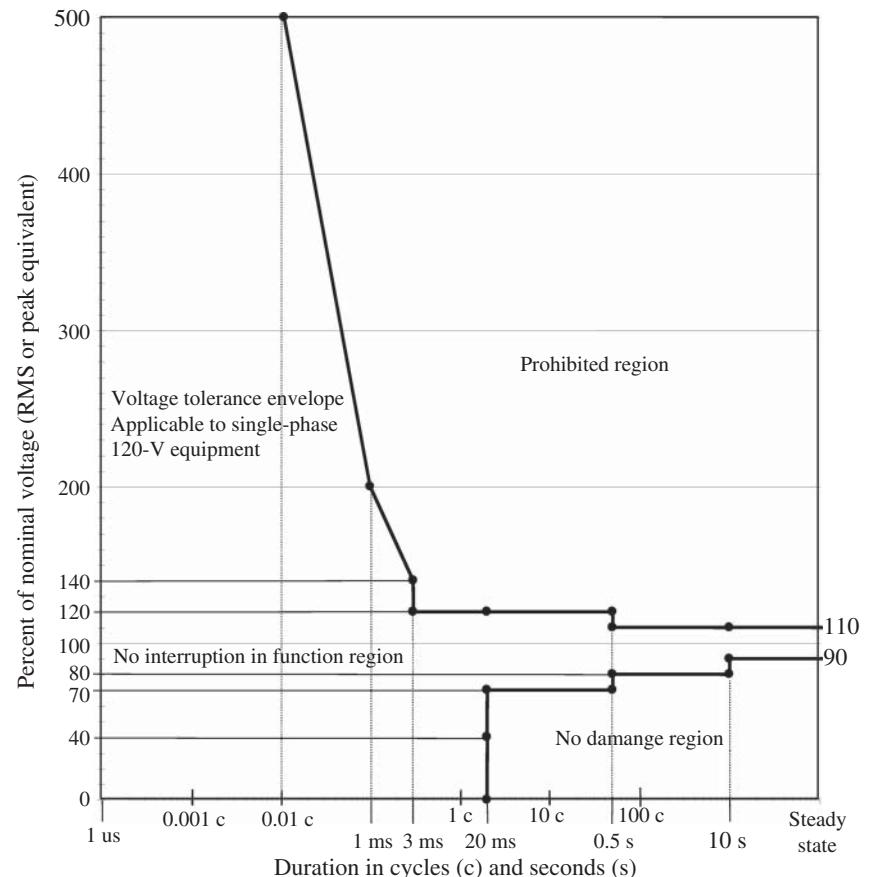


Figure 5.1 ITIC curve. Source: public domain.

electric energy consumption by reducing service voltage to customers (say, by a few percent). *Conservation voltage reduction* (CVR) as a strategy to promote energy conservation was introduced in the late 1970s and early 1980s. The effectiveness of CVR has been somewhat uneven.

Some of the difficulty in predicting the effects of voltage reduction is based on the peculiarities of motor loads, as different types of motors respond differently to voltage changes. The nonobvious effect of service voltage on various types of loads is discussed further in Chapter 6, specifically in the context of ZIP Loads (Section 6.3.1).

From an operational standpoint, where voltage is controlled by vintage distribution system hardware, there may be little room for discretion in choosing a preferred operating voltage. Many distribution operators are probably glad just to keep voltage more or less within tolerance everywhere.⁶ This situation is changing with the installation of newer voltage control technology.

Another caveat is that energy consumption is distinct from instantaneous power demand. Thermostatically controlled loads, such as heaters or refrigerators, deliver a certain amount of energy to do a specific amount of work, as determined by the end use (i.e., some number of joules

⁶ For example, a CVR program in California in the 1980s intended for utilities to narrow their voltage tolerance range from $\pm 5\%$ to $+0\%-5\%$ so as to reduce energy consumption. It could not be implemented quite as planned because of limitations in utilities' voltage-control hardware, along with the logistical challenge and cost of measurement and compliance verification.

of heating or cooling). If the power drawn by these loads is reduced due to voltage, their duty cycle will simply adapt to stay on longer, until the thermostat set point is reached. Meanwhile, the efficiency of the motor could vary with service voltage—in either direction.

Finally, overall energy consumption includes loads as well as losses on the distribution feeder itself. Suppose, for instance, that real power consumption for the aggregate load is approximately unchanged at a lower voltage, but that reactive power demand increases. In this case, line current and feeder losses would actually increase.

In sum, we would typically expect a positive correlation between service voltage and power or energy consumption, but the actual effect varies depending on circumstances, and predicting it requires detailed modeling.

5.2 Frequency

Electric grid frequency changes when generation and demand are not balanced. If demand exceeds available real power generation, energy will be drawn from the rotational kinetic energy of generators, which consequently slow down both mechanically and electrically (see Section 11.1). Conversely, in the case of overgeneration throughout the grid, frequency increases. In this context, we assume frequency to be the same everywhere throughout a *synchronous* grid.

Drifting frequency presents a risk mainly for synchronous machines, including generators and synchronous motors, as some of their windings may experience irregular current flows and become overloaded. For their own protection, synchronous generators are equipped with relays to disconnect them from the grid in the event of over- or underfrequency conditions. The sensitivity of these relays is a matter of some discretion, but would typically be on the order of 1%.

Similarly, sections of the transmission and distribution systems may be separated by over- and underfrequency relays. For example, a transmission link in a nominal 60-Hz system might have an underfrequency relay set between 58 and 59 Hz. Such a significant departure from the nominal frequency would indicate a very serious problem in the system, at which point it becomes preferable to deliberately interrupt service to some area and isolate functioning equipment, rather than risking unknown and possibly more prolonged trouble. A key objective is to prevent cascading blackouts, in which one portion of the grid that has lost its ability to maintain frequency control pulls other sections down with it as generators become unable to stabilize the frequency and eventually trip off-line. The situation is somewhat analogous to a group of mountain climbers who are roped up to support each other in case of a fall—but under some dire circumstances, it may become necessary to cut the rope.

Unlike the large frequency excursions associated with crisis events, smaller deviations can be treated by system operators with some degree of discretion. In highly industrialized areas, the choice of tolerance is driven more by cultural and regulatory norms than by the technical requirements of the grid itself. In areas facing serious supply shortages, the frequency tolerance may be much wider, as the risk of damaging equipment is weighed against the need to provide any service at all. The size of the grid is also an important factor, as it is more difficult to balance generation and demand on a small island, for example. Accordingly, there are international differences in the precision with which nominal a.c. frequency is maintained.⁷ In the major U.S. power systems,

⁷ For example, the tolerance settings for frequency deviations from the nominal 50 Hz in East and West Germany (interconnected with Eastern and Western Europe, respectively, during the Cold War) were so different that their grids could not be immediately synchronized after the 1990 reunification, and the transmission of electricity across the former border first required conversion to direct current.

frequency can be expected to fall between 59.9 and 60.1 Hz, barring any major disturbances, and usually within a much tighter range of a few hundredths of a hertz. Note that this statement applies only to large grids and not to small generators or microgrids, whose output could depart from the nominal frequency by several Hz; see also Section 11.1.4 on frequency tolerance.

One intuitive reason for maintaining a very exact frequency is that analog electric clocks (and any technology that depends on timekeeping based on the a.c. frequency) will in fact go slower if the frequency is low and faster if it is high. Grid operators in highly industrialized countries, where people and their equipment might actually care about a fraction of a second lost or gained, keep track of cycles lost during periods of underfrequency over the course of a week, and make up those cycles on a certain evening or weekend, outside regular business hours.⁸

Example

Suppose the nominal 60-Hz frequency remains low at 59.9 Hz for one entire day. How much time is “lost” on a.c. clocks?

Normally, there are 60 complete a.c. cycles in each second. The time lost corresponds to the number of cycles lost, where each cycle represents 1/60 of a second. At 0.1 Hz below nominal, 1/10 of a cycle is lost every second, or one full cycle lost every 10 seconds. This is equivalent to 60 cycles lost every 600 seconds, or one second every 10 minutes. Over the course of an hour, six seconds are lost, which comes to 144 seconds or about two and a half minutes per day. Repeating the exercise for a more realistic frequency of 59.99 Hz, the error comes to one-tenth that, or 14.4 seconds per day.

A few seconds per day is a level of inaccuracy in timekeeping that most people would hardly notice—indeed, it might compare to the error in a mechanical clockwork of average quality. It would be a gross error, however, in any advanced technological application that requires synchronization of components.⁹ But with inexpensive quartz crystals for local clocks and nearly ubiquitous access to Internet or satellite communications in the 21st century, reference to the electric grid for timekeeping has become something of a quaint anachronism.

A more important motivation for tight a.c. frequency control is the shared responsibility for balancing supply and demand, and accounting for electric power transactions across boundaries, between neighboring jurisdictions or *Balancing Authorities*. In this context, frequency errors as well as imports or exports are jointly captured in the *Area Control Error* (Section 11.1.5) as a basic coordination tool. Thus, the precision with which grid frequency is maintained is typically far more interesting to system operators than to electricity customers.

5.3 Waveform and Harmonics

A clean waveform means that the oscillation of voltage and current follow the mathematical form of a sinusoidal function. This conformance arises naturally from the geometry of the generator windings that produce the electromotive force or voltage (Section 10.1). Departures from sinusoidal waveform include both continuous and temporary, transient disturbances like the event shown in Figure 5.2. Most transient disturbances are due to transient *faults*, such as momentary contact between a power distribution line and a tree branch.

⁸ Proving that time does, in fact, go by faster on the weekends.

⁹ An insightful commentary on the preoccupation with precise timekeeping in our culture can be found in James Gleick, *Faster: The Acceleration of Just About Everything* (New York: Pantheon, 1999).

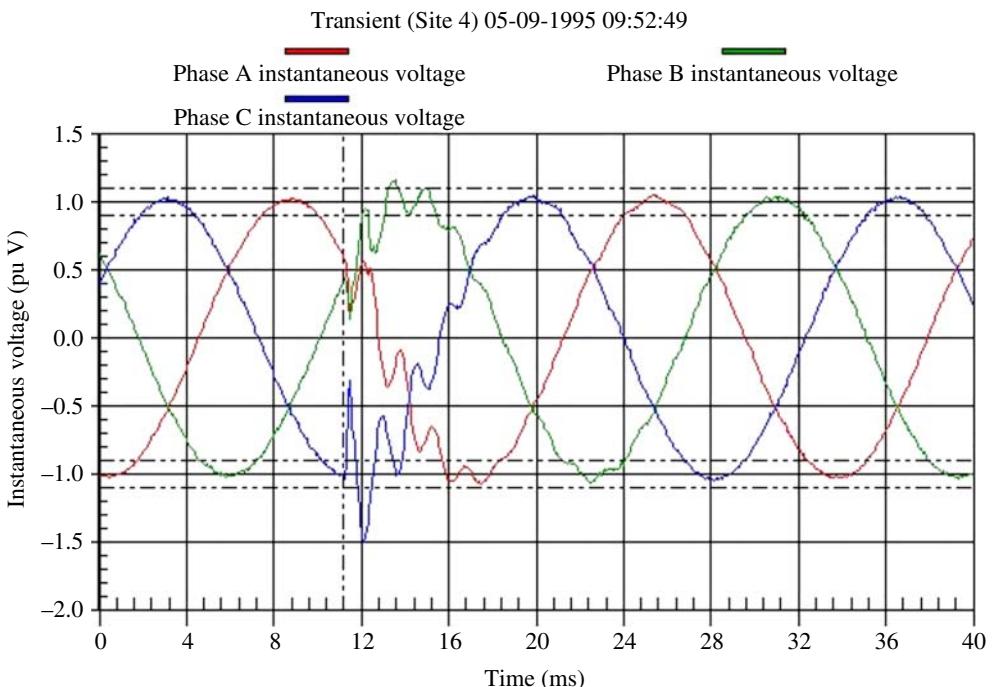


Figure 5.2 Transient voltage waveform disturbance, as seen with a PQube power quality recorder.
Source: Courtesy of Alex McEachern.

The ideal sinusoidal waveform can be altered by the imperfect behavior of all things connected to the grid: any a.c. device, whether producing or consuming power, can “inject” into the grid time variations of current and voltage, which can be observable some distance away from the offending source. As detailed in Section 5.3.1, current distortion is more common and usually of greater magnitude than voltage distortion.

Continuous distortion of voltage or current waveform is described in terms of oscillations at frequencies that are multiples of the fundamental (50 or 60 Hz) and are therefore called *harmonics*—as in music, where a harmonic note represents a multiple of a given frequency. While there are infinitely many possible higher-frequency components that can distort a wave, we care here about those that are *periodic*, that is, recurring with every cycle, and are therefore steadily observable. For this to be true, they must be integer multiples of the basic a.c. frequency, thus earning the title “harmonic.”¹⁰

When superimposed onto the basic 60-cycle wave, harmonics manifest as a jagged or squiggly appearance instead of a smooth curve. Mathematically, such a jagged periodic curve is equivalent to the sum of sinusoidal curves of different frequencies and magnitudes. The process of converting between the two representations is known as *Fourier analysis*. In Fourier analysis, a periodic function of any shape can be written as the sum of many ideal sinusoidal functions at different frequencies.¹¹ The fundamental frequency is the inverse of the period, that is, the time over which

¹⁰ Periodic disturbances can also occur at frequencies *less* than the fundamental, in which case they are called *subsynchronous oscillations*. These are more interesting in the context of wide-area grid stability (Section 13.4), rather than from a utility customer’s perspective.

¹¹ Fourier analysis is analogous to the way the human ear physically distinguishes a composite sound: by stimulating tiny hairs inside the cochlea, tuned to different specific frequencies by their length. Our brain combines

the function repeats itself, and all other components will be integer multiples of this fundamental frequency ω . In reality, a distorted voltage or current will not be exactly periodic—nor is any real voltage or current observed in the grid, considering that this would require an exactly constant frequency forever. Still, it is usually meaningful to characterize the waveform with a steady-state approximation over some time interval.

Harmonics are numbered based on which multiple of the fundamental frequency they represent. For reasons discussed below, odd-numbered harmonics tend to dominate over even-numbered harmonics. At higher frequencies (on the order of kilohertz), their contribution is usually smaller and the effect more local. Thus it is common to measure and study harmonics up to some limit, such as the 15th or the 31st harmonic.

Resistive loads are generally unaffected by waveform. Harmonics can cause vibration, buzzing, or other distortions in motors and electronic equipment, as well as losses and overheating in transformers.

In sum, a desirable waveform is one with low harmonic content, which is synonymous with a smooth, round sine wave. Significant departures from this ideal will cause heating, which generally means losses, and can create safety issues in extreme cases. Even beyond the practical and economic implications, a clean waveform entails a certain degree of engineering pride: after all, the a.c. voltage waveform is the final product delivered to the customer. Of course, everyone in the industry discovers at some point how the reality of power systems differs from our textbook abstractions—be it the waveform, balanced phases, or the behavior of generators.

Historically, there was not much utilities could do to address harmonic distortion from customer loads. With the proliferation of power electronics, harmonics are becoming more pronounced and more likely to present a problem. At the same time, it is now possible to recruit power electronics to *mitigate* harmonics on the electric grid if they are severe enough to warrant such efforts (see Section 14.4).

5.3.1 Current Versus Voltage Harmonics

The distinction between current and voltage harmonics is important. Current harmonics tend to be much more pronounced, as they are a direct consequence of *nonlinear loads*. Voltage harmonics arise as a secondary consequence of current harmonics, and depend on the *source impedance* of the grid. The term source impedance includes all impedances (such as conductors and transformers) that make up the series path from the original, ideal source (assumed to supply a perfectly steady voltage) to the load, thus characterizing the electricity source as “seen” by the load from its connection point. Specifically, the source impedance describes the extent to which the supply voltage varies with the current drawn by the load.¹²

In power systems, we generally assume that power generation sources act as voltage sources with a perfect sinusoidal waveform. When such a voltage is applied to a linear load with constant impedance, the resulting current will also be sinusoidal. With inductive or capacitive reactance, the current may be shifted in time, but the complex form of Ohm’s Law $V = IZ$ will always hold for any linear load.

Modern power electronics, by contrast, contain more than just passive impedances. Any device with semiconductor components such as diodes and transistors or any internal on-off

the intensity of signal at each frequency (the power spectrum) into a composite that we interpret as a recognizable sound with a distinctive timbre, like a person’s voice. The same note played on a different musical instrument has the same fundamental frequency but a different harmonic spectrum.

¹² The effect of source impedance is the same physical phenomenon described in a different context as *voltage regulation*; see Section 8.6.

switches—but also an analog device that becomes saturated as its internal electric or magnetic field reaches a limit—will present itself to the circuit as a different impedance during different portions of the cycle. For example, if a small capacitor inside a device is fully charged up during a fraction of a cycle, it will take a “gulp” of current and then stop. The instantaneous current, therefore, is nonlinear with respect to the instantaneous voltage applied.

For the most part, though, the direction of the current will remain consistent with the direction of the applied voltage, even in a nonlinear load where it is distorted in shape.¹³ Moreover, the distortion is usually the same in the positive and negative directions, as most devices are indifferent to the sign of the voltage and current (e.g., the saturation limit of a transformer core is the same with either polarity), and thus the distortion appears as opposite and symmetrical gulps or dips in the waveform. In Fourier analysis, it can be seen that those harmonics that contribute to a waveform with opposite distortion are the odd-numbered harmonics.

As a consequence of a nonlinear current drawn by the load, the voltage as measured at the load will also suffer some distortion, due to the voltage drop between the (ideal) generation source and the load, which is the product of current and source impedance. A sample nonlinear current and its effect on voltage, mediated by the source impedance, is illustrated in Figure 5.3. The inductive part of the source impedance is a function of frequency (since $X = \omega L$) and increases for the higher harmonics; this subtlety must be taken into account for accurate analyses of voltage distortion.

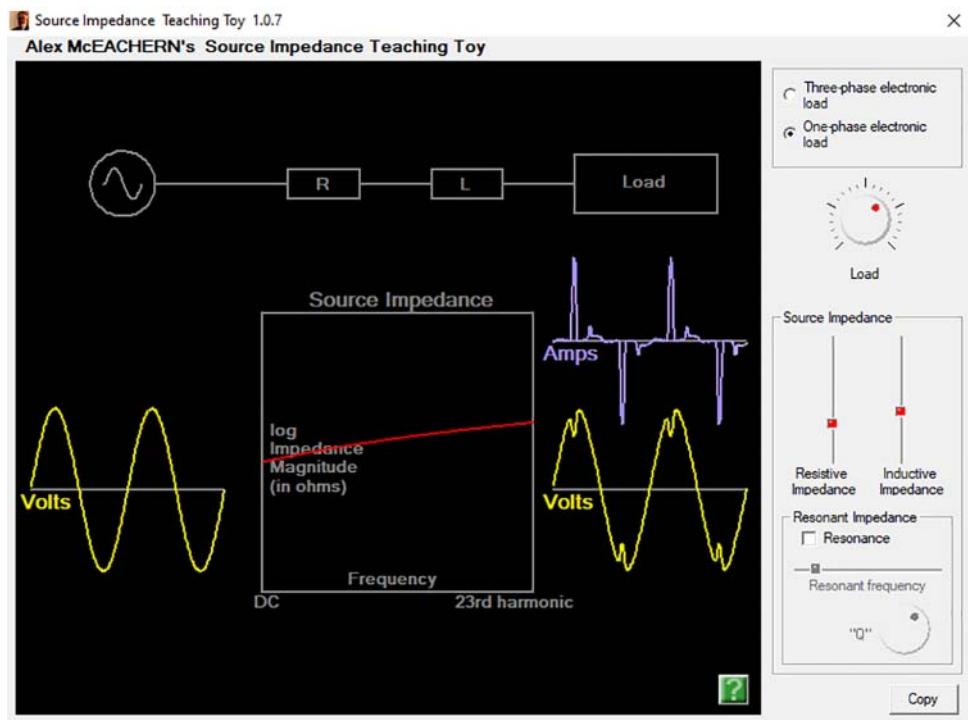


Figure 5.3 Simulated current drawn by a highly nonlinear load and its effect on voltage due to source impedance, visualized in the Power Quality Teaching Toy.

¹³ This makes sense because the device acts as a load—whereas, to the extent that current and voltage point in opposite directions, a device generates rather than consumes power.

As a general rule of thumb, one would expect the source impedance throughout a network to cause a voltage drop on the order of several percent (say, 5%) between zero load and full rated load at any given location. Therefore, even a large “gulp” of current tends to cause only a small notch in the voltage.

Although smaller in magnitude than current harmonics, voltage harmonics are more consequential, because they affect neighboring loads on nearby circuits. When a highly nonlinear load distorts the current, this current must travel only through the delivery infrastructure (i.e., conductors and transformers) to the load, where it causes unwanted heating, but it would likely go unnoticed by other loads connected in parallel. To the extent that the service voltage is distorted, however, this directly impacts all other loads connected at that location, which are now seeing a nonsinusoidal voltage waveform. Assuming they are linear loads, their own current will then become distorted in proportion with the voltage supply.

5.3.2 Quantifying Harmonic Distortion

The relative contribution of higher-frequency harmonics compared to the fundamental frequency can be quantified as *harmonic content* or *total harmonic distortion* (THD). THD is expressed as a percentage indicating the amount of power carried by the harmonic frequencies, as opposed to the power contained in the fundamental 50- or 60-hertz wave. A common standard for power generation equipment such as inverters is to produce voltage THD below 5%. Current harmonics can include much higher distortion levels (see Figure 5.3) due to loads that cannot easily be identified or regulated.

Note that some distortions can be temporary or transient (that is, occurring only briefly). In that case, it makes little sense to attempt to describe the waveform in terms of harmonic content. Reference to THD generally assumes that the condition being described is ongoing and steady.

Total harmonic distortion of current THD_I is quantified in terms of a sum of current contributions from all harmonic frequencies $k\omega$, where the summation index k goes from 2 to infinity. The analogous definition applies to voltage distortion, THD_V . Because power depends on the square of current, the contributions of harmonic currents at different frequencies are weighted based on I^2 , and then normalized by taking the square root to produce a meaningful ratio to the fundamental current (denoted with subscript 1). It doesn’t matter if one uses the amplitudes or rms values for each, as long as they are consistent within the ratio. Expressed in percent by convention, THD_I is defined by the equation

$$\text{THD}_I = \frac{\sqrt{\sum_{k=2}^{\infty} I_k^2}}{I_1} \cdot 100\% = \frac{\sqrt{\sum_{k=2}^{\infty} I_{k,\text{rms}}^2}}{I_{1,\text{rms}}} \cdot 100\% \quad (5.1)$$

By rearranging, we can express the total rms current as the geometric sum of the fundamental and harmonic currents:¹⁴

$$I_{\text{rms}} = I_{1,\text{rms}} \sqrt{1 + (\text{THD}_I/100)^2} \quad (5.2)$$

In sum, the total current, which includes all the contributions from various harmonics, is greater than the fundamental (50- or 60-Hz) current alone.

¹⁴ Recall that root-mean-square quantities are defined for any periodic function (Section 3.1.3), so that writing V_{rms} or I_{rms} does not imply any assumption about sinusoidal waveforms. I_{rms} in Eq. (5.2) refers to the distorted, nonsinusoidal waveform, while $I_{1,\text{rms}}$ and $I_{k,\text{rms}}$ in Eq. (5.1) refer to the ideal sinusoidal constituents of the mathematically decomposed waveform.

Harmonic currents deliver zero real power to the load when multiplied by the fundamental voltage. Like a current shifted by 90°, the harmonic current at some multiple of the voltage frequency will yield a positive or negative instantaneous power product in equal amounts, averaging to zero regardless of its phase shift:

$$[V \cos(\omega t) \cdot I \cos(n\omega t + \phi)]_{\text{ave}} = 0$$

Since voltage waveform is usually quite close to a pure sinusoid at the fundamental frequency, it is generally true that only the fundamental current contributes useful work. Harmonic currents still produce heat as they travel through equipment. Like losses at the fundamental frequency, losses due to harmonic currents are captured by I^2R . Thus, the current I used to calculate resistive heating should be the root-mean-square current magnitude that includes all frequencies. The ability of harmonic currents to cause resistive heating throughout devices and conductors implies losses and can also pose a safety hazard—thus motivating standards for limiting permissible THD.

5.3.3 Distortion Power Factor

The current contribution from harmonics is analogous to the component of current that is displaced or shifted in time relative to the voltage, in that neither serves to transfer net power, yet both contribute to losses. This reality motivates an extended definition of power factor. Power factor as defined in Chapter 3 was limited only to *displacement power factor*.

In that context, we defined $p.f.\cdot_{\text{disp}} = \cos \theta$ in terms of the phase angle difference θ between current and voltage.

More generally, power factor can be defined as the ratio of average to apparent power:

$$p.f. = \frac{P}{S}$$

Conventionally, these two definitions are assumed to agree with each other, but realistically, in the presence of nonideal waveforms, they do not agree.¹⁵ The total apparent power $S = I_{\text{rms}} V_{\text{rms}}$ in the denominator may include circulating power not only due to inductive or capacitive reactance that causes a steady time shift (displacement) between current and voltage, but also due to nonlinear loads that produce a mismatch of waveforms. Note that for any component of current that is not time-aligned with voltage—whether due to a phase shift or difference in frequency—power transfer averages to zero. Thus, the current associated with harmonics contributes to overall current magnitude and apparent power S , but not real power P .

Separating the two effects of displacement and distortion, one can define a *distortion power factor*, $p.f.\cdot_{\text{dist}}$:

$$p.f.\cdot_{\text{dist}} = \frac{1}{\sqrt{1 + (\text{THD}_1/100)^2}} \quad (5.3)$$

¹⁵ Like reactive power, “power factor” is not a physical quantity with an inherent existence apart from a measurement process agreed upon by convention. There is no intrinsically “correct” way to define or measure reactive power or power factor, and this topic has been quite controversial historically. At issue is whether the measurement process, and the implicit model on which it is based, is usefully related to reality. See H. Kirkham, A. Emanuel, M. Albu, and D. Laverty, “Resolving the Reactive Power Question,” *IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, Auckland, New Zealand, 2019. DOI: 10.1109/I2MTC.2019.8826915.

The overall power factor,¹⁶ which accounts for displacement as well as distortion effects, can be written as the product of the two:

$$p.f. = p.f.\cdot\text{disp} \cdot p.f.\cdot\text{dist} = \frac{P_{1,\text{ave}}}{V_{1,\text{rms}} I_{1,\text{rms}}} \cdot \frac{1}{\sqrt{1 + (\text{THD}_1/100)^2}} \quad (5.4)$$

where $V_{1,\text{rms}}$ and $I_{1,\text{rms}}$ refer to the fundamental component only.

As power electronic devices of all sizes are becoming ubiquitous on the grid and often replacing inductive loads, distortion power factor is becoming an increasingly relevant metric. More so than customer loads, transmission lines and transformers with their inductive properties remain responsible for lagging displacement power factor on the system, which necessitates reactive compensation (Section 3.4.5). There is little to be done about distortion power factor by the electric utility in today's standard practice, although it is possible to actively cancel offending harmonics with power electronic devices designed for this purpose.

5.3.4 Transformers and Triplen Harmonics

When considering harmonic effects, location is important. Recall that alternating currents of different frequencies behave differently as they pass through electrical equipment because reactance (inductive or capacitive) is frequency-dependent. Since the grid is generally dominated by inductive behavior from transformers and transmission lines, it presents a higher impedance to higher frequency signals (acting as a low-pass filter). Consequently, harmonics in the voltage and current waveforms are *attenuated* over electrical distance in the grid, and are mostly considered a local phenomenon.

A single transformer will tend to filter a majority of current and associated voltage harmonics generated by nonlinear loads, so that one would not expect to see significant effects on the other side of the transformer from the offending device. In the process, waste heat is produced inside the transformer windings as well as in the magnetic core, which resists rapid reversals. The result is that the transformer operates at lower efficiency, and may face a shortened life span due to chronic overheating. At the same time, transformers themselves can also contribute harmonics. This happens whenever the magnetic core saturates, and an increasing current no longer corresponds to an increasing magnetization (see Section 8.5.1). In this way, the transformer presents a nonlinear load to the source.

Certain harmonics exhibit special behavior due to symmetries. In particular, any harmonic that is a multiple of three has the special property of coinciding for all three phases. Consider the third harmonic of the a.c. base frequency: in a 60-Hz system, this means a small oscillation at 180 Hz. As illustrated in Figure 5.4, this third harmonic of Phase A is indistinguishable from that of Phase B or C, since the functions are just superimposed on each other. The same is true for all multiples of the third harmonic (6th, 9th, 12th, 15th, etc.), called *triplen* harmonics. Since odd-numbered harmonics tend to dominate over even-numbered ones, and since there is diminished power in the higher harmonics, the triplen harmonics of greatest practical relevance are the 3rd and 9th.

One important consequence of the fact that triplen harmonics coincide for each phase is that there is no voltage difference between any pair of phases. But for a delta-connected transformer, the voltage on each of its three primary windings is the voltage *difference* between a pair of phases (A-B, B-C, or C-A). Therefore, no voltage at any triplen harmonic frequency is detected by a delta-connected device, and no power is transferred.

¹⁶ This is often called the *true power factor*, but in view of the previous footnote, it is perhaps best to eschew the word "true."

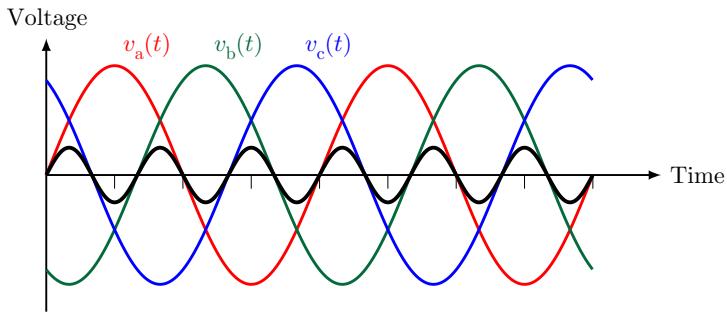


Figure 5.4 The third harmonic of all three phases coincides.

For a wye-connected device, the triplen harmonics are additive. This means that the voltage at the neutral point will vary relative to ground, and a current will flow in the neutral conductor that is the sum of the triplen harmonic components of all three phases.

Neutral conductors are often designed with the expectation that they only need to carry a small current, much less than the rated phase current. This is because aside from harmonics, the current in the neutral conductor would be nominally zero. In reality, it would amount to the phase imbalance, which should be a small percentage of the phase current. If the neutral current due to triplen harmonics is greater than accounted for in the neutral conductor size, this can present a safety issue due to overheating.

Note that one-third of all the integer multiples of the base frequency are also multiples of three. Therefore, roughly one-third of all the power contained in harmonic components of a wave is blocked by any delta connection, and combined in the neutral of any wye connection. Triplen harmonics are also described as *zero-sequence harmonics*, because they share the key property of the zero-sequence component (see Section 4.2) of the fundamental 60-Hz waveform in an imbalanced three-phase system: namely that all three (ABC) phases coincide.¹⁷

Incidentally, this property also applies to any components of direct current, which can be induced especially on long transmission lines by external phenomena such as solar storms. Geomagnetically induced d.c. currents that are equal on all three phases become additive in the neutral conductor, to potentially cause thermal damage to transformers. It is possible to block such currents with a series capacitance in the neutral, at some expense and design trade-offs.

Problems and Questions

- 5.1** Referring to the ITIC curve in Figure 5.1, state what you expect would likely happen to a typical load if it experienced the following voltage excursions:
- 130% of nominal voltage for a duration of 5 milliseconds (ms).
 - 60% of nominal voltage for a duration of 50 ms.
 - 80% of nominal voltage for a duration of 0.2 seconds.
- 5.2** It has been a strange day for grid operators. At 11:00:00 A.M., the synchronous frequency (nominally 50 Hz) dropped to 49.900 Hz and stayed there for exactly 55 minutes. The system

¹⁷ Mathematically, all the integer harmonics in a three-phase system can be seen as corresponding to symmetrical components in staggered groups depending on their modulus 3, with the 1st, 4th, 7th, 10th, etc. harmonic being positive-sequence and the 2nd, 5th, 8th, 11th, etc. harmonic being negative-sequence.

then operated at a frequency of 50.100 Hz for exactly five minutes. What time is shown by a synchronous wall clock (perfectly set prior to the event) at 12:00:00 noon?

- 5.3** Find the amount of total harmonic distortion in percent that would produce a distortion power factor of 0.8.
- 5.4** In your own words, explain why harmonic distortion is considered a “local” phenomenon in the electric grid.
- 5.5** Consider a load connected to a 60-Hz sinusoidal voltage source: $v(t) = 169.7 \cos(2\pi 60.0t) = \sqrt{2} 120 \cos(\omega t)$ V. The load is nonlinear and draws a current given by $i(t) = \sqrt{2} 12 \cos(\omega t - 30^\circ) + \sqrt{2} 6 \cos(3\omega t - 10^\circ) + \sqrt{2} 3 \cos(5\omega t)$ A.
- What is the real power transferred to the load?
 - Find the total harmonic distortion (THD) for current, in percent. Does the phase shift of the harmonics matter?
 - What are the displacement, distortion, and true power factor?
 - Try your hand at a rough sketch of the current waveform.
- 5.6** A single-phase load draws an rms current 20 A at 240 V and delivers 3.36 kW of real power.
- Based on this information, what is the power factor?
 - Without giving too much thought to the exact nature of the load, the owner decides to install a parallel capacitor for reactive compensation adjacent to the load. The capacitor’s kVAR rating is intended to correct the power factor to 1.0, under the assumption that we are dealing solely with a displacement power factor. What is that kVAR rating? Draw a diagram illustrating the presumptive P , Q , and S for the load and capacitor.
 - What is the current in the capacitor under these expected conditions?
 - What is the new expected total current for the load–capacitor combination?
 - The current for the load–capacitor combination is measured after the capacitor is installed. Surprisingly, the ammeter now reads 18 A (rms). How might you explain this situation? Draw a new diagram based on your hypothesis.
 - If your hypothesis is correct, what would have been the right kVAR rating for the capacitor to achieve a unity displacement power factor?
 - What would be the expected total current for this latter load–capacitor combination?

6

Loads

6.1 Types of Loads

In the context of electric circuits, the term *load* refers to any device in which power is being dissipated (i.e., consumed). Physically, we think of load in terms of the electrical characteristics of individual circuit elements. From the circuit perspective, a load is defined by its *impedance*, which comprises a resistance and a reactance.

The impedance of an individual device may be fixed, as in the case of a simple light bulb, or it may vary in time. In practice, almost all things connected to the grid are internally complicated and contain large numbers of diverse electrical components, which collectively present themselves as some aggregate impedance at the plug in the wall outlet. This impedance may change slowly (for example, if an appliance has several operating settings), or within each a.c. cycle (in the case of *nonlinear* loads).

In the larger context of power systems, loads are further aggregated: rather than describing an individual appliance, “load” may refer to an entire household, a city block, or all the customers within a city or state. For the power industry, “load” therefore has attributes beyond impedance that relate to aggregate behavior, such as the timing of peak demand or the ramp rate over the course of an afternoon. Aggregation thus occurs on multiple levels, and the term “load” can apply to vastly different scales, from circuit boards to regional power markets.

If we consider a load as being defined by its impedance, there are theoretically three types of loads: purely resistive loads, inductive loads, and capacitive loads. *Resistive loads* are those consisting basically of a heated conductor, whether a heating element in a toaster oven or a glowing filament in a light bulb. *Inductive loads* are the most common and include all types of motors, fluorescent lights, and transformers like those used in power supplies for lower-voltage appliances—basically, anything with a coil in it. *Capacitors*, by contrast, do not lend themselves for doing mechanical or other practical work outside electrical circuitry (arc welding machines are one notable exception). While capacitors are standard components of electronic circuits and capacitance occurs *within* many appliances, it does not tend to dominate their overall electrical appearance to the power system. Thus, capacitive loads are uncommon on the macroscale.

Power electronics are an important and rapidly growing type of load in the 21st century. Unless one happens to be versed in the design details of a particular appliance (which may be proprietary to the manufacturer), it is nearly impossible to predict from first principles how a given electronic device will appear to the grid when it is plugged in (say, capacitive or inductive), or how it will behave under different circumstances: one just has to measure it to find out.

Table 6.1 Examples of different types of loads.

Pure Resistive Loads	Incandescent Lamps
Motors	Heaters (oven, toaster, iron, radiant space heater) Compressors (air conditioner, refrigerator, heat pump) Pumps (well, pool) Fans Household appliances (washer, mixer, vacuum cleaner) Power tools
Electronics	Large commercial three-phase motors (grocery store chiller) Power supplies (computers, televisions, small adapters) Battery chargers (electric vehicles, electronic devices) Microwave ovens Electronic motors (linear compressor) Fluorescent and LED lamps

Loads differ in the type of electric power they can use. A standard informational label should state the nominal voltage or acceptable range, frequency, and the rated current.¹ Most motors are designed for alternating current at a specific frequency and voltage, although there are also direct-current motors. Pure resistors are the most forgiving loads, being tolerant of low voltage and completely indifferent to the direction of current flow; any resistor will operate interchangeably on a.c. or d.c. Electronic devices use direct current internally, but are manufactured to interface with the standard a.c. grid voltage through a power supply. Many common appliances are sold in DC versions for use in camper vehicles or remote homes. Table 6.1 shows some illustrative examples of different types of loads.

In this chapter, we begin with the characteristics of the most common loads at the device level: resistive loads, motors, and electronic equipment. Section 6.2 deals with the physical connection between loads and the distribution system, Section 6.3 examines load response to voltage in more detail, and Section 6.4 considers load from a system operator's perspective.

6.1.1 Resistive Loads

The simplest type of load is a purely resistive load, that is, one without capacitive or inductive reactance. Many familiar appliances fall into this category: incandescent light bulbs and all kinds of resistive heaters, from toasters to electric blankets, space heaters, and electric ranges. In each case, the heating element consists simply of a conductor that dissipates power according to the relationship $P = I^2R$, which can be rewritten as $P = IV$ or $P = V^2/R$ by substituting Ohm's law.

As can be recognized from $P = V^2/R$, the amount of power dissipated increases with decreasing resistance, given that a constant voltage is supplied. Therefore, the strongest heaters are those with the lowest resistance. Although a given device would be best characterized by its resistance in ohms, this information is not usually stated on the package, as most users would have no idea how to

¹ To indicate that appliances are compatible with different voltage or frequency standards, a slash is used, such as 120/240 VAC or 50/60 Hz. A tilde, as in 100~490 V, indicates that the device can operate anywhere within that continuous range.

interpret it. What people care about is how much power, in watts, a device consumes. To specify the power rating, it is necessary to assume a voltage at which this device is to be operated; this assumption should be stated on the label along with the power rating.

In general, resistive loads are the simplest to operate and the most tolerant of variations in power quality, meaning variations in the voltage level above or below the nominal 120 V, or departures of a.c. frequency from the nominal 50 or 60 Hz. A resistive heating element can be damaged by excessively high voltage, which will cause it to overheat or wear out prematurely. The filaments of incandescent light bulbs, for example, will burn out sooner if they are operated at a higher voltage. Low voltage, on the other hand, causes no physical damage whatsoever in a resistive device, though the heat or light output will be reduced. Resistive loads are essentially indifferent to the frequency of a.c., or whether it alternates at all. For example, the performance of an incandescent lamp at 120-V_{a.c.} and 120-V_{d.c.} would be indistinguishable to the human eye.

As for voltage magnitude, despite the efforts of utilities to maintain it at a constant level, the actual voltage as seen by an appliance plugged into an outlet varies from one time and location to another. The tolerance for voltage supplied by U.S. utilities is typically $\pm 5\%$, which corresponds to an actual range of 114–126 V based on a nominal 120 V for residential customers (see Section 5.1). What happens to loads and their power consumption when voltage changes? Unlike motor loads, resistive loads are easily predictable in this regard. It follows from the equation $P = V^2/R$ that if the voltage is increased, the power drawn by resistive loads increases as the square of the voltage.

Example

Compare the power consumption of a 100-W purely resistive load at 114 V versus 126 V.

We may assume that the load was rated for a nominal voltage of 120 V; thus, its resistance is $(120 \text{ V})^2/100 \text{ W} = 144 \Omega$. At 114 V, this bulb draws a power of $(114 \text{ V})^2/144 \Omega = 90.25 \text{ W}$. At 126 V, it draws $(126 \text{ V})^2/144 \Omega = 110.25 \text{ W}$.

Another way to approach this calculation is to note that voltage is being increased or decreased from nominal 120 V by 5%, meaning that it is changed by a factor of 1.05 or 0.95, respectively. To find the change in the amount of power, we can simply square these factors, since power for a purely resistive load is proportional to voltage squared. A 5% voltage increase thus multiplies power by 1.1025, and a 5% voltage decrease multiplies power by 0.9025. These factors, when multiplied by 100 W, give the same results as above.

6.1.2 Dimmer Circuits

Voltage response is also relevant to dimmer circuits, where the brightness of a light is controlled by way of reducing the effective root-mean-square (rms) voltage across the lamp. One type of dimmer switch contains solid-state circuitry that “clips” the voltage at a maximum value below its sinusoidal peak. Another type of dimmer circuit uses *pulse-width modulation* and turns the voltage on and off rapidly for some adjustable average fraction of the cycle. Such circuits are increasingly common with electronic loads, including d.c. circuits. For example, even battery-powered light-emitting diodes (LEDs) can have a noticeable flicker or strobe effect, revealing that their direct-current supply is in fact being chopped by a switching circuit so as to reduce the average power.

Because resistive loads are quite indifferent to waveform or harmonics, it doesn’t matter if electronic dimmers produce a rather jagged voltage. Incandescent lamps will radiate instantaneous power depending on the temperature of the filament, which has some thermal inertia; the human eye can’t discern any flicker at 50 or 60 Hz. What we do see is that the overall brightness is diminished in relation to the average amount of power dissipated by the filament, which in turn

is given by the average or rms voltage over the course of each cycle. This average can be visualized in terms of the area left under a clipped or chopped-up voltage sine wave.

Electronic loads may or may not be compatible with dimmer switches. For example, many but not all LED lamps are designed to tolerate modified voltage signals with a reduced rms voltage; compatible ones may be labeled as *dimmable*. Most fluorescent lamps are not dimmable, as they interact with the a.c. power supply by way of a *ballast* and power converter that produces a high-voltage, high-frequency signal and is already complicated enough. In general, reducing the voltage for appliances other than simple resistors—especially motors—could damage them, or they may not work at all. Dimmer switches should therefore only be installed in appropriate circuits.

For purposes of reviewing basic circuit analysis from Chapter 2, it is instructive to consider a simple, old-fashioned alternative to electronic dimming: Can't the voltage be reduced by just inserting another resistance in series with the load? Yes, and a variable resistance used for this purpose is called a *rheostat*. The problem with the rheostat is its own power dissipation—that is, heat. This implies not only waste, but a possible risk of overheating.

As illustrated by the following example, a rheostat's series resistance is added to the resistance already present in a light bulb, which reduces the total current (because it must now flow through both resistances; see Section 2.2.1). Because of the lower current, the light bulb will dissipate less power and appear dimmer. Another way to view the situation is that the supply voltage (say, 120 V) is now split between the light bulb and the rheostat according to the relative proportion of their resistances. The greater the resistance of the rheostat, the greater a fraction of the supply voltage it will sustain, meaning a lower voltage and less power for the light bulb. At the same time, a significant amount of power may be dissipated by the rheostat, which could easily get too hot in a confined space. By contrast, the waste heat generated by an electronic dimmer is usually insignificant both in terms of safety and losses.

Example

A 100-W incandescent light bulb in a 120-V circuit is dimmed to half its power output using an old-fashioned rheostat, or variable resistor, as a dimmer. What is the value of the resistance in the dimmer at this setting, and how much power is being dissipated by the rheostat itself?

First, let us determine the resistance of the light bulb by considering its normal operating condition, when the power is 100 W at 120 V. Rewriting $P = V^2/R$ as $R = V^2/P$, we obtain $R = (120 \text{ V})^2/100 = 14,400/100 = 144 \Omega$. We are given the information that with the dimmer in series, the power drops from 100 to 50 W. In this situation, we do not know the voltage drop across the light, because now $120 \text{ V} = V_{\text{light}} + V_{\text{dimmer}}$ according to KVL (Section 2.3.1). But since we know the light's resistance and its power at the new operating condition, we can determine the current using $P = I^2R$, which gives $I = (50 \text{ W}/144 \Omega)^{1/2} = 0.59 \text{ A}$ (for comparison, the current at 100 W was 0.83 A).

Knowing the current, we can now infer the total resistance in the circuit including the dimmer by using Ohm's law $V = IR$, or $R = V/I = 120 \text{ V}/0.59 \text{ A} = 204 \Omega$. Thus, the rheostat's resistance at this setting is $R_{\text{dim}} = 204 - 144 = 60 \Omega$.

The power dissipated within the rheostat is $P = I^2R = (0.59 \text{ A})^2 \cdot 60 \Omega = 21 \text{ W}$. The light–dimmer combination consumes $50 + 21 = 71 \text{ W}$. This is less than the 100-W light by itself, but considerably more than the 50 W we might have expected if we failed to consider what happens inside the dimmer. Note that continually dissipating even a modest 21 W inside an electrical box in the wall could pose a fire hazard.

When $R_{\text{dim}} \approx 0$, the rheostat has no effect, acting simply as a piece of conducting wire that, ideally, dissipates no power at all. As R_{dim} gets very large, the current becomes very small, until the light

goes out. The fraction of power dissipated within the rheostat increases with higher resistance setting, although the total power (switch plus light) decreases.

6.1.3 Motors

Electric motors represent a substantial fraction of residential, commercial, and industrial loads.² Motor loads comprise fans, pumps of all kinds including refrigerators, air conditioners and heat pumps,³ power tools, and anything else electric that moves. Electric transportation is often powered indirectly through battery chargers, as in the case of cars and trucks, or through a dedicated a.c. or d.c. infrastructure that serves electric rail. These motor loads are ‘seen’ by the electric grid as the power converters at the interface.

A motor is essentially the same thing as a generator operated backward; electrical and mechanical energy are converted into one another by means of a magnetic field that interacts with both the rotating part of the machine and the electrons inside the conductor windings. The mechanical power output of a motor is conventionally expressed in units of *horsepower* (hp), where 1 hp = 0.746 kW, only to distinguish it from the electrical power expressed in kilowatts. The essential physical properties of motors are analogous to those of generators discussed in Chapter 10.

Aside from differences in size and power, there are three distinct types of motors that correspond to the three main types of generators: induction, synchronous, and d.c. In each case, the motor is similar to its generator counterpart. Induction motors are the least expensive and remain by far the most common, especially for smaller applications.

Like the induction generator, the induction motor’s rotational speed varies with the *torque*⁴ applied to the rotor, and thus the amount of power transferred. In order to produce a torque by magnetic force, the induction motor requires *slip*, or a difference between the motor’s mechanical rotation speed and the a.c. frequency. The rotor magnetic field comes from an induced current in the rotor, which has no independent electrical source, but receives electromagnetic induction from the stator windings (supplied by the a.c. source) as a result of the motor’s internal geometry.⁵ When an induction motor is started from rest, the rotor does not yet have any current circulating inside it, and thus no magnetic field; the a.c. current supplied to the stator windings thus encounters very little impedance for the first fraction of a second. The resulting phenomenon characteristic of

2 A 2011 study estimated that electric motor-driven systems account for between 43% and 46% of all global electricity consumption. P. Waide and C.U. Brunner, *Energy-Efficiency Policy Opportunities for Electric Motor-Driven Systems* (International Energy Agency, May 2011).

3 Heat pumps are increasingly popular for space or water heating because of their greater *coefficient of performance* (COP) as compared to resistive heating (which is already 100% efficient at converting electrical into thermal energy). A heat pump essentially refrigerates the ambient air and rejects the heat into a compartment where air or water is intentionally warmed up. Since this heat is already present and only needs to be moved “uphill” thermodynamically from cold to hot, the amount of heat transported can be several times greater than the mechanical work done in the process. The COP is the ratio of heat delivered to work done. COPs of 3 or 4 are readily achieved for moderate temperature gradients, and a heat pump uses correspondingly less electricity than a resistive heater. For more about heat engines, see also Section 15.1.2.

4 Torque implies the application of force to rotate something. Technically, it is the product of force (in units of pounds or newtons) and the distance from the rotational axis (in feet or meters) where the force is applied; thus, the units of *foot-pounds* are used for torque. As we know from direct experience, it is easier (requiring less force) to turn something (apply a certain torque) by pushing at a point farther away from the center of rotation—say, turning a nut with a longer wrench, or using a screwdriver with a thicker handle. Oddly, in terms of physical dimension, torque corresponds to energy. This makes sense when we note that to produce units of power, we multiply torque by the rotational or angular frequency, which has units of inverse time. Thus, power equals torque times revolutions per minute (rpm).

5 The rotor is the rotating part and the stator the stationary part of the machine; see Section 10.3.

induction motors is called the *inrush current*. The inrush current accounts for the familiar flicker of lights when a heavy motor load is starting up, as the local line voltage is momentarily reduced by the voltage drop associated with the high current flow.

Even after the rotor magnetic field has been established, any motor consumes additional power and current as it mechanically accelerates to its operating speed; this is the *starting current*. A typical motor's starting current may be five to seven times greater than the current under full load. Larger and more sophisticated commercial motor systems include starting controls designed to soften the impact of motor loads on the local electrical system while gradually ramping up their rotational speed.

Because of the geometric relationship between the rotor and stator and the need to induce a rotor current, an induction machine always consumes reactive power (VAR; see Section 3.4). This is true regardless of whether the machine is operating as a motor or generator (i.e., consuming or generating real power). Therefore, induction motors all have power factors less than unity; depending on the motor design and rating, typical power factors range from below 60 to the low 90s of percent. Induction motors are chiefly responsible for lagging power factors in most systems.

Synchronous motors, by contrast, have an independent source of magnetization for their rotor, which may be either a permanent magnet or an electromagnet created by an external current (*excitation*). This independent magnetization allows the machine to operate at synchronous a.c. speed—some fraction of 3600 rpm for 60 Hz, depending on the number of magnetic poles—regardless of load. Like a synchronous generator, it is possible to operate a synchronous motor at different power factors. Synchronous motors are more complicated and expensive than induction motors of comparable size. They are characteristically used in industrial applications, especially those requiring high horsepower and constant speed.

D.C. motors also have independent magnetization. Most important, they require *commutation* of the d.c. in order to produce the correct torque. Commutator rings or brushes involve moving electrical contacts that are inherently prone to mechanical wear; d.c. machines therefore tend to require more maintenance than a.c. machines. D.C. motors are distinguished by a high starting torque, which is useful in applications such as accelerating vehicles from rest. They also afford convenient speed control, since the rotational speed varies directly with voltage.⁶ While many smaller d.c. motors are used in off-grid applications, the most important d.c. motor loads in power systems throughout the last century have been electric trains and streetcars. Other types of motors include hybrids, such as the switched reluctance motor (a variation on the d.c. motor with electronic commutation). Variable speed motor drives, which present to the grid as power electronic loads, can offer substantial efficiency gains and are increasingly common.

Besides motor type, another important distinction is between single- and three-phase motors. Like a generator, a motor benefits from the constant torque afforded by three separate windings, staggered in space and time, that in combination produce a rotating magnetic field of constant strength (this applies only to a.c., not d.c. machines). Three-phase motors therefore operate more smoothly and much more efficiently than single-phase motors, though they are also more expensive. They are commonly used for large industrial and commercial applications where high performance, including high horsepower output and high efficiency, is essential, and where three-phase utility service is standard. Most smaller commercial and almost all residential customers in the U.S. receive only single-phase service because of the significant cost difference in distribution.⁷

6 A classic example is a model electric train, where the locomotive speed can be controlled quite precisely with a simple rheostat dial that varies the voltage between the tracks.

7 The standard combined 120/240-V service, though it affords different options for plugging in appliances, is still a single phase; see Section 6.2.

Accordingly, there is not a significant market for smaller three-phase motors to date. It is technically possible, however, to operate a three-phase motor on single-phase service by inserting a (rotary or electronic) *phase converter* that effectively splits the voltage and current along several circuits and changes their relative timing to produce three staggered sine waves.

Because motors account for such a large portion of our society's energy consumption, they represent a significant opportunity for efficiency improvements. Typical energy conversion efficiencies range from an average of 65% for small (fractional horsepower) to over 95% for very large (over 500 hp) motors. The operating efficiency depends both on the motor design and the operating condition, with most motors running most efficiently near their rated capacity (the load for which they were presumably designed). A variety of approaches exist to increase motor efficiency, ranging from specific motor designs to systemic aspects such as speed controls and motor sizing in relation to the load. Short of replacing a motor with a more efficient model, some of the standard options include rewinding existing motors with conductors of lower resistance and adding adjustable or variable speed drives (known as ASD or VSD) to allow for energy savings when less power is required.⁸

Today's standard ASDs control speed by rectifying the a.c. supply to d.c. and inverting it again to a.c. of variable frequency. The associated conversion losses are easily outweighed by the significant savings in mechanical energy. When a fan or pump motor moves a fluid (air or water),⁹ the volume moved and the kinetic energy imparted to the fluid both increase with motor speed (volume linearly and energy with the square of velocity), making the power theoretically proportional to the cube of rotational speed. Friction may have an additional impact. A doubling of motor speed thus implies roughly an eightfold increase in power, or, in practical terms, a speed reduction of 20% can yield 50% energy savings.

It is interesting to note that motors can have a very long life, on the order of 100,000 operating hours. Depending on how heavily a motor is used, the lifetime cost of supplying its energy can be much greater than the motor's initial capital cost; in fact, it is not unusual for a commercial motor's annual electric bill to exceed its purchase price by an order of magnitude. Therefore, the additional cost of a more efficient motor system may be recovered in a reasonable time even if the percentage efficiency gain appears small.

Unlike resistive loads, electric motors are sensitive to power quality, including voltage, frequency, harmonic content and, in the case of three-phase machines, phase imbalance. One of the key problems that tend to afflict motors is unequal and excessive heating of the windings, which leads to energy and performance losses, degradation of the insulating material, and possibly short-circuiting. Such heating can be caused by voltage levels that are either too high, too low, or too uneven across phases. Voltage also affects the power factor at which a motor operates. Excess heating and energy losses can further result from harmonic content in the a.c. sine wave, in addition to undesirable vibration. Finally, transient disturbances (brief spikes or notches) in voltage may impact motor controls, the commutation mechanism, or protective circuit breakers. Because of these sensitivities, it is not uncommon for owners of expensive and sophisticated motor systems to install their own protective or conditioning equipment, so as to guarantee power quality beyond the standard provided by the local utility.

⁸ Nadel et al. (*Energy-Efficient Motor Systems*) offer an excellent review of strategies for increasing motor efficiency.

⁹ This most common type of load is called a *variable torque load*, as opposed to *constant torque loads* (where torque and speed are independent) and *constant power loads* (where torque and speed are inversely related). The categorization alone illustrates some of the subtlety and complexity of electric motor systems.

6.1.4 Electronic Devices

Consumer electronics—basically, anything that has little buttons on it—are powered by low-voltage direct current (d.c.). They may be operated either by batteries or through a *power supply* that delivers lower-voltage d.c. by way of a step-down transformer and rectifier. Power supplies may be external and plug directly into the outlet with an adapter, or they may be contained within the larger appliance (e.g., a computer). Electronic loads, once a minor curiosity, are on their way to becoming a dominant load category. This is due not only to their proliferation in number, but the fact that electronic power supplies are increasingly associated with larger power equipment.

It is important to distinguish between devices that do real “work” in the physical or chemical sense (such as heating or moving objects, or charging a battery) and electronics whose job is merely to move information, encoded by way of tiny circuits switched on or off in some pattern. The real work is usually done by a d.c. load behind the power converter. Moving information does not in and of itself require physical work, except for the inadvertent heating of circuit elements when current flows through them. The power consumption of any electronic appliance can be gauged by how warm it gets during operation. This type of heat is very much like the waste heat resulting from mechanical friction, which can be reduced by clever design but never completely eliminated.

To the extent that the job of an electronic circuit component is only to relay the information whether a particular circuit is “on” or “off,” this job can be performed with a very small current. Indeed, an important aspect of the continuing evolution of electronics, from vacuum tubes to transistors to integrated circuit chips, has been to reduce the operating currents to ever smaller amounts while fitting them into ever smaller spaces without overheating.

“Pure” electronic devices use very little energy, as evidenced by the fact that they can operate for a long time on a small battery—if not on a few square centimeters of solar cells in indoor light. Then there are those appliances whose external power supplies send low-voltage d.c. through a rather skinny electrical cord, which tells us that the power consumption had better be small. Most electronic devices require a modicum of physical work in addition to their “brains,” such as lighting up a screen, moving a loudspeaker membrane, or emitting electromagnetic radiation to communicate.

Even when electronic appliances require little physical work to operate, they also contribute to energy use by way of *standby power*, which is consumed whenever the machine is plugged in and ready to respond. For example, any device that can be switched on by remote control needs to keep some internal circuits activated to enable it to recognize the remote signal. This standby power can sometimes be recognized as warmth on the back of the appliance. Similarly, power supplies that remain plugged into a live outlet continually dissipate some finite amount of heat. Though it doesn’t sound like much, a constant drain of 4 W, for example, adds up to 35 kWh of waste heat over the course of a year. For a household full of electronic gadgets, an annual consumption of hundreds of kilowatt-hours for standby appliances is not unusual. A simple power strip with a switch allows energy-conscious consumers to avoid this phenomenon without too much hassle.

Many conventional electronic loads are in fact combinations of information-processing circuits and something that does relevant physical work. A microwave oven, for example, may have an impressive array of buttons, beeps, and displays, but its real job is to deliver hundreds of watts in the form of electromagnetic radiation (which it does by running a current through coils at high frequency), to be picked up by resonating water and fat molecules in our food. Television and computer screens are shooting photons at the viewer. Most computers need a fan for cooling their central processing unit (thus the sound of the computer being “on”), and a laser printer uses the bulk of its electrical energy to heat the drum. Again, it is the amount of heating that gives us a

direct measure of the electric power consumption involved, as all of the energy entering an appliance from the outlet must eventually exit and be dissipated as heat.¹⁰ An important special case is the battery charger (see Section 6.1.4), a d.c. load whose job it is to move ions “uphill” against a chemical potential. The work done is converted later (into useful things and still, ultimately, heat) when the battery is discharged.

While mixed electronic and power appliances may appear in various proportions as resistive, inductive, or capacitive to the power system, their most important characteristic is that they are *nonlinear* loads which can impact power quality. Because electronics contain nonlinear circuit elements, they tend to produce harmonic distortion of the current (Section 5.3.1) and are primarily characterized by a *distortion power factor* (Section 5.3.3) rather than *displacement power factor* (Section 3.4.2).

Finally, loads vary in their own sensitivity to power quality. For example, the proliferation of LED clock displays in the late 20th century (prior to manufacturers’ standard practice of including small batteries) had the unintended consequence of vastly increasing consumer awareness of momentary grid voltage fluctuations: some readers may recall the iconic blinking 12:00 on the video cassette recorder.

6.1.5 Electric Vehicles

The major new development in electric loads is electric vehicle (EV) charging. An EV charger is essentially a rectifier (Section 14.4) that provides d.c. to drive a chemical reaction in the car battery (Section 15.3.2). The power electronics for the a.c.-to-d.c. conversion may reside in the separate charger unit, onboard the vehicle, or both. Unlike small electronics, an EV charger’s function is to do real physical work, and a lot of it. With a typical battery capacity on the order of tens of kWh (and some greater than 100 kWh), an EV can easily double the electric load of an entire household, both in terms of instantaneous power demand, and cumulative energy usage.

In U.S. nomenclature, *Level 1* chargers plug into common outlets at 120 V. At a maximum current of 20 A, this corresponds to 2.4 kW of power. *Level 2* chargers use 240 V on a dedicated circuit. For example, a Level 2 charger with a maximum current of 48 A constitutes a peak load of 11.5 kW. *Level 3* refers to commercial fast-charging stations that take higher service voltages from the utility and deliver up to hundreds of kilowatts.

At the time of this writing, EV charging demand is poised to grow dramatically in the transition toward carbon-neutral transportation. Beyond adding to overall electric energy use, EV load growth may trigger the need to upgrade local capacity (e.g., service transformers) and challenge voltage regulation on distribution circuits (Section 7.4). An opportunity to mitigate these challenges lies in the eminent controllability of battery chargers. On a time scale of minutes to hours, many users don’t actually care about when, exactly, the battery is charging or at what rate. It stands to reason that this flexibility could be leveraged with appropriate economic incentives and controls (see Section 6.4.4). Power electronics also have the intrinsic ability to affect power quality—for better or for worse, again depending on the control systems that are required or incentivized.

¹⁰ Especially in commercial environments, this waste heat is often masked by aggressive air conditioning, not only for human comfort but to protect the machines themselves from overheating (e.g., computer servers). In this sense, heat-producing electronics represent a double energy load. Efficiency improvements to such equipment, including “energy-saver” features for turning devices off when not in use, thus afford the double benefit of reduced air-conditioning needs.

6.2 Single- and Multiphase Connections

Taking full advantage of three-phase transmission, certain loads connect to all three phases. These are almost exclusively large motors, such as those in heavy machinery or commercial ventilation and refrigeration equipment, where efficiency gains from smooth three-phase operation are worth the extra cost. Three-phase motors contain three windings that make up three distinct but balanced circuits. A three-phase machine plugs into its appropriate power outlet with three phase terminals and one neutral to handle any current resulting from phase imbalance. However, most utility customers in the United States and many other countries do not have three-phase service. All the familiar loads from residential and small commercial settings represent a single circuit with only two terminals to connect.

The U.S. standard, 120-V nominal socket thus has two terminals (Figure 6.1), a “hot” or *phase* (black wire, small slot) and a *neutral* (white wire, large slot), in addition to a safety *ground* (bare or yellow-green wire, round hole). Ungrounded outlets that predate electrical-code revisions can sometimes still be found older buildings.

The phase supplies an alternating voltage with a nominal rms value of 120 V between it and the neutral terminal. The neutral terminal is ostensibly at 0 V, but its voltage will tend to float in the range of a few volts or so, depending on how well the nearby loads are balanced among the three phases, and on the distance to the point where the neutral terminal is physically grounded.

The differently sized slots in the receptacle accommodate the polarized plugs that are standard for lighting appliances. The purpose is to reduce the risk of shock for people changing a light bulb with the circuit energized: the small prong delivers the phase voltage to the back of the fixture where it is less likely to be touched accidentally.¹¹

The ground is not part of the power circuit except during malfunction. It should connect to the earth nearby—typically, through a building’s water pipes—and serves to protect against shock and fire hazards from appliances due to faulty wiring, exposure to water, or other untoward events.¹²

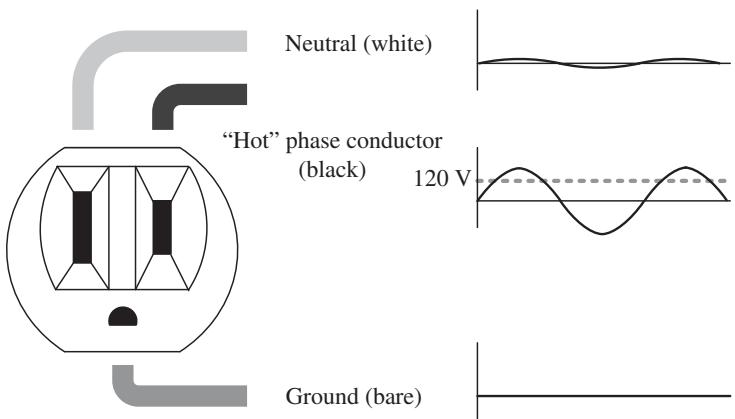


Figure 6.1 Standard U.S. electrical outlet or wall socket.

11 This paragraph seems to beg for some screw-in-a-lightbulb joke, but I can’t quite put my finger on it.

12 While Figure 6.1 illustrates the installation of the socket in the common orientation that is somewhat reminiscent of a face, it is technically preferable to orient it with the ground on top. This is in consideration of a failure mode where a plug is partially inserted, exposing the energized metal prongs, and the possibility of a metal item (falling from above) contacting both phase and neutral to cause a fault. If only the ground prong is touched, nothing happens.

Removing the ground prong from a plug to fit it into an ungrounded outlet is not just contrary to the advice of appliance manufacturers, but probably a bad idea. The appropriate way to use an adapter plug made for ungrounded outlets is to fashion an external ground with a wire connecting to the small metal loop on the plug.

Most utility customers also have wiring for higher voltage appliances, which may provide 240 or 208 V. This is why the utility service enters the house with three wires, which are not the same as the three phases. Instead, they are one neutral and two phase conductors.

There are two approaches for delivering different voltage levels to a single building. The *split phase* design is standard for residential service in North America, and delivers a 120/240-V combination from a single phase (A, B, or C). In this case, the two phase conductors tap the same winding of a distribution transformer at different points. The transformer has the correct turns ratio so that the full secondary coil provides 240 V. By tapping the secondary coil at the halfway point, another wire can supply half the voltage, or 120 V. Figure 6.2 illustrates the situation.

Because the primary and secondary circuits are linked only magnetically, not electrically, and because voltage is inherently a relative, not an absolute quantity, the neutral wire on the secondary side can be connected to any arbitrary part of the winding, forcing that point of the winding to be at or near ground potential while the voltage of other points is simply measured relative to this neutral terminal.

For example, if the neutral connects to one end of the secondary winding, the winding can be tapped in the middle for 120 V and at the other end for 240 V.

If the neutral connects to the center tap, as in Figure 6.2, either end of the winding will yield 120 V, with opposite polarity. This split-phase arrangement allows for more efficient utilization of the three wires, because each hot or phase conductor can serve any of the customer's 120-V circuits. What the individual 120-V load does not see, though, is that these two sets of 120 V are measured

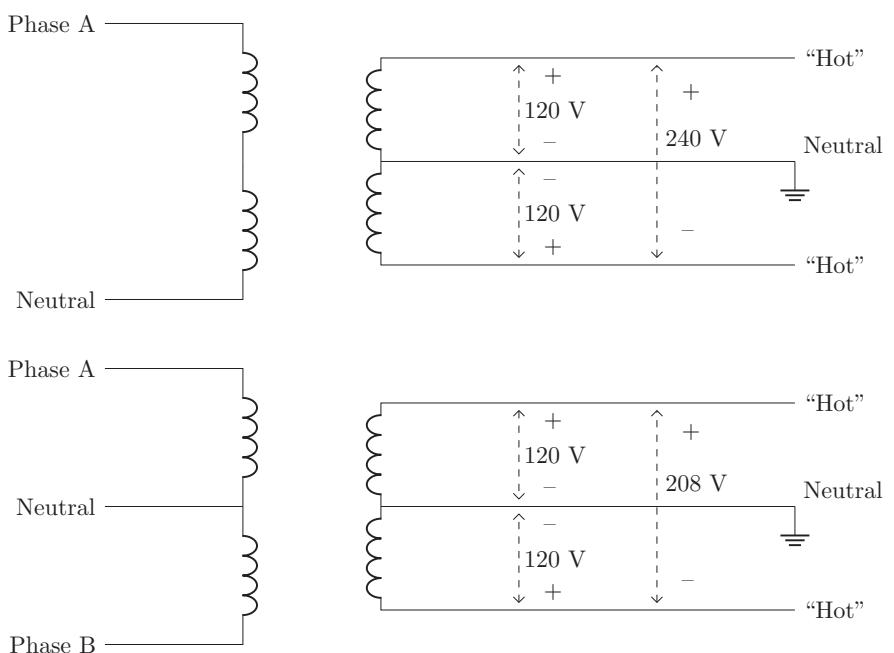


Figure 6.2 Transformer taps and multiphase service.

in opposite directions—in other words, they are 180° apart in phase, making them additive. Thus a load connected between these two (rather than the neutral) will see their sum, or 240 V.

An alternative design is the 120/208 V supply, where two different phase combinations are tapped. The 120 V corresponds to the *phase-to-neutral* or *line-to-neutral* voltage between one phase (say, A) and the neutral terminal, while the 208 V corresponds to the *phase-to-phase* or *line-to-line* voltage between two different phases (say, A and B).¹³ As described in Section 4.1.4 and illustrated in Figure 4.4, this phase-to-phase voltage exceeds the phase-to-neutral voltage by a factor of $\sqrt{3}$, or about 1.732, in a way that is intuitive only to connoisseurs of trigonometry. Note that $208 \approx 120\sqrt{3}$. Many but not all loads designed for a nominal 220 or 240 V supply can function satisfactorily on 208 V.

When three phases are provided to a building, individual circuits may supply various single- or three-phase loads. A single-phase circuit could provide either the line-to-line or the line-to-neutral voltage. Three-phase circuits will terminate in special three-phase receptacles supplied by all three phases and neutral (plus a ground for safety) for three-phase loads. The three-phase load may be internally wired in a wye or delta configuration (see Section 4.1.4), meaning that it can avail itself of three sets of line-to-neutral voltages or three sets of line-to-line voltages.

Clearly, it is important to ensure compatibility between loads and voltages available at the outlet. This goes for the voltage range supplied as well as the current rating. Though a.c. devices can be damaged by attempting to operate on too low a voltage, the most dangerous combinations would be a higher voltage supplied than what the load was designed for (dissipating more power than expected, and perhaps causing some insulation to fail), or a higher current draw than what the receptacle or the wires behind it were designed for. Standardized sockets of different shapes are intended to prevent such unfortunate combinations, even across different countries, and prevent the unwitting user from accidentally plugging an appliance into the wrong outlet, or in the wrong orientation. Many different standards for plugs and outlets of various shapes exist in different countries.¹⁴

6.3 Voltage Response of Loads

6.3.1 ZIP Load Model

From the standpoint of the utility, load is traditionally not considered controllable, and subject to the whims of the customers. However, load may also vary in response to the exact voltage supplied, in ways that are not obvious.

For the purpose of estimating voltages and power flows on distribution circuits, we need a mathematical model for the relationship between the voltage supplied and the real and reactive power drawn by the load. The most straightforward case, which we have assumed in all the introductory chapters, is that any given load has a constant impedance. This makes sense from a purely physical

¹³ The distinction between neutral and ground is a very important one in the context of abnormal operating conditions (e.g., when a large current might flow from neutral to ground), or when making precision measurements where a small nonzero voltage on the neutral conductor matters. For purposes of a basic overview as in Figure 6.2, ground and neutral can be considered equivalent.

¹⁴ Standards are a vital topic in power systems, where the compatibility among many diverse components is essential for everyone's safety. A standard—that is, an agreed-upon way of doing things—is called for whenever two conditions apply: first, there is more than one “correct” way of doing things, and second, it is important that everyone do things the same way. Standard plugs and sockets used in the United States can be found at https://en.wikipedia.org/wiki/NEMA_connector (accessed February 2024).

standpoint: If a load just consists of pieces of metal with some particular fixed shape, then it has a resistance and reactance that is, for all practical purposes, fixed. Therefore, a standard assumption would be that load impedance is constant, and power is given by the familiar equations

$$P = \frac{V^2}{R} \quad \text{and} \quad Q = \frac{V^2}{X}$$

The quadratic voltage dependence is important. It means that for a given percentage change in voltage magnitude, there will be a more pronounced change in power demand. For example, if voltage is increased by 1%, under constant impedance we would expect the power to increase by 2%.

In reality, loads consist of more than just simple, fixed pieces of metal. Not only are there moving parts, but independent internal controls that continually change the effective impedance the load presents to the circuit. When there are power electronic circuit boards in between the main power supply and the physical task at hand—lighting up a screen, playing music over a speaker, tumbling a load of laundry, microwaving a plate of food—it's anyone's guess how the instantaneous current will actually vary relative to the instantaneous voltage. For anything but the simplest circuits, it would be next to impossible to predict the actual functional dependence of current or power on voltage from first principles. In practice, then, the power–voltage relationship for loads has to be determined empirically, from measurements.

Such measurements will generally yield a scatter plot of observed real or reactive power versus voltage, for either an individual load or an aggregation of any size. It is then possible to fit a curve to the data. There are many different ways to parameterize such a curve. Typically, it will not be smooth, and it can vary dramatically not just across different types of devices, but even among different manufacturers and models of otherwise similar appliances.

The most standard approach is to identify three mathematical components: a quadratic, a linear, and a constant term. This amounts to representing the load as a composite of three superimposed contributions: a load with constant impedance, one with constant current, and one with constant power. In reference to the variables Z , I , and P , this load model is called “ZIP loads.”

There is no reason to assume that the voltage dependence of real and reactive power would be the same for any given load, since P and Q would be governed by different physical behaviors within the load. Therefore, we need a separate fitting for real and reactive power.

In equation form, we express the real and reactive power P and Q at an operating voltage V relative to the power P_0 and Q_0 at a reference voltage V_0 . We then have six parameters or coefficients to estimate: Z_p , I_p , P_p , Z_q , I_q , and P_q .¹⁵ These coefficients are dimensionless numbers (positive or negative) that add to 1 so as to preserve the overall scale.

$$P = P_0 \left[Z_p \left(\frac{V}{V_0} \right)^2 + I_p \left(\frac{V}{V_0} \right) + P_p \right] \quad (6.1)$$

$$Q = Q_0 \left[Z_q \left(\frac{V}{V_0} \right)^2 + I_q \left(\frac{V}{V_0} \right) + P_q \right] \quad (6.2)$$

Equations 6.1 and 6.2 describe relationships between load behaviors under two different operating conditions, V and V_0 , which allows us to make a statement like, “If the operating voltage is increased or decreased by some percentage, the real and reactive power will increase or decrease by some other percentage.” It is understood that this quantitative relationship will not hold over an arbitrarily wide range of voltages, but usually we are interested in realistic operating ranges within, say, $\pm 10\%$ of nominal.

¹⁵ There are other formats in the literature, but the number of parameters will still be six.

There is not always a good physical intuition for these relationships. An incandescent lamp or resistive heater would be expected to behave mostly as a constant-impedance load. By contrast, electric motors could have all sorts of control systems; for example, the objective may be to deliver a constant amount of mechanical power. If a motor operates less efficiently at a lower voltage, it is possible that the electrical power demand will actually increase in order to keep mechanical power the same when the voltage decreases.

Example

Suppose a refrigerator has the following ZIP coefficients:

$$Z_p = 1.17, \quad I_p = -1.83, \quad P_p = 1.66, \quad Z_q = 7.07, \quad I_q = -10.94, \quad P_q = 4.87$$

If this load draws 200 W and 200 VAR at 120 V, what are its real and reactive power at 114 V and at 126 V? How does the power factor change with voltage? Plugging the given quantities into Eqs. (6.1) and (6.2), we have:

$$\begin{aligned} P_{126} &= 200 \text{ W} \left[1.17 \left(\frac{126}{120} \right)^2 - 1.83 \frac{126}{120} + 1.66 \right] \\ &= 200 \text{ W}(1.29 - 1.92 + 1.66) = 200 \text{ W}(1.03) = 206 \text{ W} \end{aligned}$$

$$\begin{aligned} Q_{126} &= 200 \text{ VAR} \left[7.07 \left(\frac{126}{120} \right)^2 - 10.94 \frac{126}{120} + 4.87 \right] \\ &= 200 \text{ VAR}(7.79 - 11.49 + 4.87) = 200 \text{ VAR}(1.18) = 236 \text{ VAR} \end{aligned}$$

$$\begin{aligned} P_{114} &= 200 \text{ W} \left[1.17 \left(\frac{114}{120} \right)^2 - 1.83 \frac{114}{120} + 1.66 \right] \\ &= 200 \text{ W}(1.06 - 1.74 + 1.66) = 200 \text{ W}(0.98) = 195 \text{ W} \end{aligned}$$

$$\begin{aligned} Q_{114} &= 200 \text{ VAR} \left[7.07 \left(\frac{114}{120} \right)^2 - 10.94 \frac{114}{120} + 4.87 \right] \\ &= 200 \text{ VAR}(6.38 - 10.39 + 4.87) = 200 \text{ VAR}(0.86) = 172 \text{ VAR} \end{aligned}$$

The power factor varies from

$$p.f_{120} = 0.707$$

$$p.f_{126} = \cos \left(\tan^{-1} \frac{236}{206} \right) = 0.66$$

$$p.f_{114} = \cos \left(\tan^{-1} \frac{172}{195} \right) = 0.75$$

We see that for this particular load, both real and reactive power have a positive correlation with voltage, but Q is more sensitive. Thus, the power factor decreases with increasing voltage, and increases at lower voltage. Note that this result is neither intuitive nor generalizable, since we don't know any details about the compressor motor or its controls.

Different variants of the ZIP load model exist that use different parameter definitions. For example, only one set of constant-impedance, constant-current, and constant-power coefficients may be used, but with a separate set of three power factor coefficients to capture both real and reactive power response. In any case, it will always take six parameters to describe load behavior.

Like power factor, ZIP models apply to individual as well as aggregate loads at any scale. A common question concerns the ZIP characteristics for an entire distribution feeder. In the context of *conservation voltage reduction* (CVR, see Section 5.1.1), the question is, how much energy could be saved by narrowing the operating voltage band to reduce the average service voltage? Clearly, the answer depends on how the actual loads on a given circuit respond to changes in voltage, as described by the ZIP model above.

6.3.2 Transient Response

We may also be interested in the load behavior during voltage sags—that is, voltages momentarily dropping to anywhere between zero and nominal. Here we are in the realm of transient, not steady-state conditions. The key question in this context does not concern the exact power demand at a low voltage, but simply whether the load can “ride through” and stay online, or whether it will disconnect. While standards exist to articulate expected behavior (such as the ITIC curve in Figure 5.1), the actual response of any given device to voltage sags is hard to predict and can vary significantly among types of appliances, makes, and models. Also, the response may depend on the detailed shape and duration of a voltage excursion.

One special case of interest concerns induction motors *stalling* at low voltages (see Section 10.6). Induction motors are common in residential and small commercial air conditioning compressors, which constitute a significant fraction of loads, especially in cooling-dominated climates and on hot days. When the supply voltage drops below some threshold, the motor lacks the torque necessary to overcome the mechanical resistance it is pushing against. But as it slows down and eventually stops turning, current continues to flow through the motor windings. These windings, without the rotation that makes them transfer mechanical power, present an almost purely inductive load that draws a large lagging current.

Such currents may themselves exacerbate and prolong a low-voltage condition. In the phenomenon called *fault-induced delayed voltage recovery* (FIDVR), a voltage sag on a distribution circuit due to a nearby fault—a fairly commonplace occurrence that is usually expected to resolve within a fraction of a second, as protective devices clear the fault—can leave the voltage depressed for minutes, if many air conditioners stall. The situation is resolved by disconnect switches within the load that respond either to thermal overload (from I^2R heating of the windings) after some time, or to an intelligent sensor that detects the stall condition. As air conditioners drop offline, the feeder voltage recovers.

In the meantime, voltage regulation devices (see Section 7.4) on the circuit may have responded to the low-voltage condition: for example, some capacitor banks automatically switch in with a short time delay upon detecting a voltage below their threshold setting. As loads disconnect, this can then lead to a temporary overvoltage and even some back-and-forth, if capacitors switch off again before the loads return.

As more types of devices such as solar and battery inverters or actively controllable loads are recruited for managing voltage on distribution systems, it is important to coordinate voltage regulation devices by choose their control settings so as to avoid oscillation behavior or “hunting.” At the time of this writing, the capabilities exist for inverters and electronically controlled loads like electric vehicle chargers to manage their real and reactive power in response to prices, communication signals or direct measurements of line voltage (and frequency, or even voltage phase angle) in order to help stabilize conditions on the circuit. The actual implementation or systematic recruitment of such controls is still nascent, however. Power electronic interfaces for generation, storage, or loads can, in principle, be programmed to do just about anything within

their physical capacity; the challenge is figuring out what to tell them to do. Robust control algorithms will be informed by a thorough understanding of the interactions among different components across the entire network—the subject of ongoing research in power systems.

6.4 Load in Aggregate

From the perspective of the grid, individual customers and their appliances are small, numerous, and hardly discernible as distinct loads. This section deals with *aggregate* load, that is, the combined effect of many customers both in terms of the magnitude and timing of electric demand.

6.4.1 Historical Context

While consumers typically think of their electricity usage in terms of a quantity of *energy* (in kilowatt-hours) consumed over the course of a billing period, the quantity of interest to system operators and planners is the *power* (in kilowatts or megawatts, measuring the instantaneous rate of energy flow) demanded at any given time. The term *demand* thus refers to a physical quantity of power, not energy. Serving that instantaneous demand under diverse circumstances is the central challenge in designing and operating power systems, and the one that calls for the majority of investment and effort.

It is interesting to reflect upon the historical service philosophy that considers demand as the independent variable that is to be met by supply at any costs. The assumption embedded in both the hardware design and the operating culture of electric power systems is that customers freely determine how much power they want, and that it is the job of power system designers and operators to bend over backwards as needed to accommodate this demand. In the same vein, power engineering texts refer to “load” as an externally given quantity, a variable beyond control, in a completely un-self-conscious manner. This assumption is codified in the social compact between utilities as regulated monopolies, and the public: in essence, the utility is given the privilege of being the exclusive service provider within a geographic territory, in exchange for assuming the obligation to serve just about any load that may exist or grow within that territory.

More recently, with power system economics scrutinized and reconsidered from different angles, the philosophy by which demand rules the game has been challenged in some respects. We now expect customers to vary their demand according to electricity prices, which are in turn driven by supply. Section 6.4.4 discusses efforts to make demand more responsive, including scenarios with remote-controlled and automated devices. The profound nature of this conceptual shift cannot be overstated; it is probably the most significant change to the mission of electric power systems since their inception. How the transition from a service-driven to a market- and data-driven system will eventually play out remains to be seen. As of this writing, operating and planning decisions in power systems still consider demand largely as an independent variable.

Short of aiming to control demand, an important discipline is dedicated to *forecasting* demand at any given future time. Drawing on detailed statistics about past behavior, meteorology, and any other factors from school holidays to new technologies that might impact electricity use, accurate load forecasting on various time scales—from minutes, hours and days to years—is both a science and an art. One growing challenge for these models, besides net demand varying with the contributions from renewable resources, is the increasing uncertainty range in weather forecasts due to climate change.

6.4.2 Coincident and Noncoincident Demand

In the utility context, analysts distinguish *coincident* and *noncoincident* demand. Coincident demand refers to the amount of combined power demand that could normally be expected from a given set of customers, say, a residential block on one distribution feeder. By contrast, the noncoincident demand is the total power that would be drawn by these customers if all their appliances were operating at the same time. It is called noncoincident because all these demands do not usually coincide. Coincident demand reflects the statistical expectation regarding how much of these individual demands will actually overlap at any one time.

For example, suppose each of 10 residences had a 600-W refrigerator. The noncoincident demand associated with these refrigerators would be 6,000 W. Under ordinary circumstances, however, the compressor in each refrigerator goes on and off on a duty cycle, and so is operating only part of the time, let's say 20%. The duty cycles of these 10 refrigerators will usually be at random in relation to each other, so we could expect that at any given time, only one in five is operating. The coincident demand in this case would be 20% of 6000 W, or 1200 W. Clearly, this sort of statistical prediction becomes more reliable when greater numbers of customers are involved.

Although ordinarily the utility observes only the coincident demand, it must be prepared to face noncoincident demand under certain circumstances. Suppose, for example, that there is an outage that lasts for a sufficient time period—an hour or so—to let all the refrigerator compartments warm up above their thermostat settings. Now power is restored. What happens? All 10 compressors will kick in simultaneously, and the 6000-W noncoincident demand suddenly coincides!

What makes this particular scenario most troublesome, actually, is not just the simultaneous operation of normal loads; rather, it is the split-second inrush current of electric motors as they turn on and establish their internal magnetic field. The sum of these inrush currents from refrigeration and air-conditioning units can overload distribution transformers and even cause them to explode the moment that power is restored after an outage. For this reason utilities often request their customers to switch most appliances off during an outage until they know the service is back.

6.4.3 Load Profiles and Load Duration Curve

Instantaneous demand, as it varies over the course of a day, is represented in a *load profile*. A load profile may be drawn at any level of aggregation: for an individual electricity user, a distribution feeder, or an entire grid. It may represent an actual day, or a statistical average over typical days in a given month or season. The maximum demand, which tends to be of greatest interest to the service provider, is termed the *peak load*, *peak demand*, or simply the *peak*.

From the power system perspective, it is sometimes relevant to compare periods of higher and lower demand over the course of a year. Thus, one might compile the highest demand for each month and plot these 12 points, indicating the seasonal as opposed to the diurnal rhythm. In warmer climates where air conditioning dominates electric usage, demand will tend to be *summer-peaking*; conversely, heating-dominated regions will see *winter-peaking* demand.

A different way to represent a load profile for this purpose is by way of a *load duration curve*. The load duration curve still depicts instantaneous demand at various times (generally in one-hour intervals), except that the hours are sorted not in temporal sequence, but ranked according to the demand in each hour. Thus, the highest demand hour of the year appears to be the first hour, followed by the second highest demand hour (which may well have occurred on a different day), and so on. Each of the 8760 hours of the year then appears somewhere on the graph, with the night hours mostly at the low-demand end on the right-hand side.

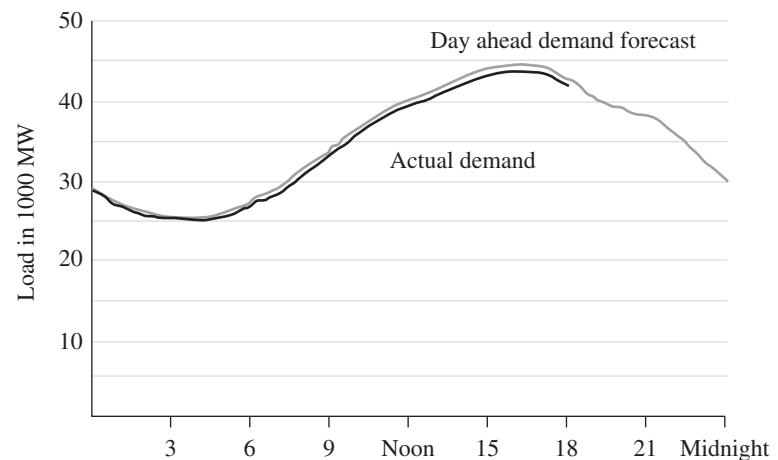


Figure 6.3 Example of a load profile from California ISO. Source: Adapted from California Independent System Operator.

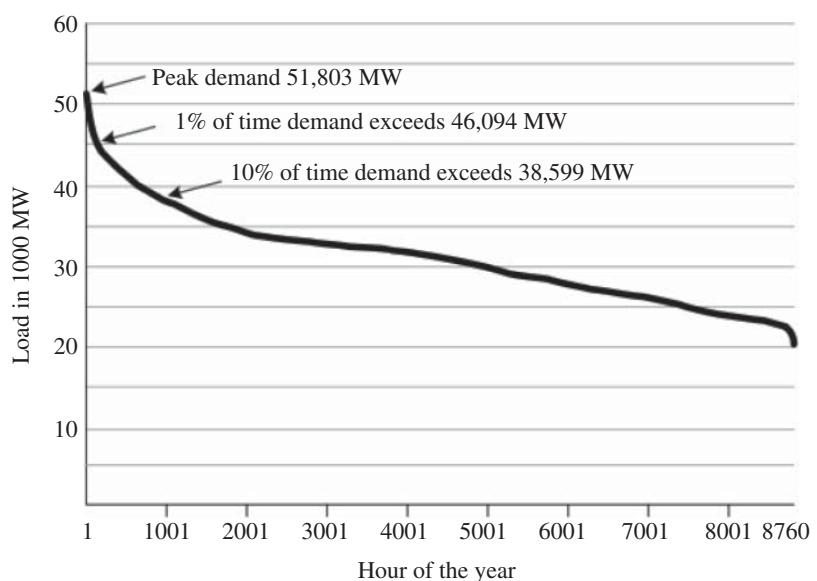


Figure 6.4 Example of a load duration curve. Source: Adapted from California Energy Commission/State of California.

Figures 6.3 and 6.4 illustrate a daily load profile and a load duration curve for the state of California from the early 2000s.¹⁶ The vertical scale is the same in both graphs, but the time axis has a different meaning in each. If the data were for the same year, we might cross-reference the graphs as follows: at 5 P.M. on August 5 (a warm but not excruciatingly hot day), the load was about

¹⁶ The daily load profile is published live by the California Independent System Operator (CAISO), along with the load forecast for the day and available generation resources (not shown here). While this particular day turned out to be slightly cooler than expected, the close proximity of forecast and actual demand offer a glimpse into the sophistication of load forecasting.

43,000 MW. By inspection of the load duration curve, we see that this would put 5 p.m. August 5 within about the top 250 or 300 demand hours of the year.

The shape of the load duration curve's peak, which is obvious at a glance, is a useful way to characterize the pattern of demand. Quantitatively, the ratio of average to peak demand is defined as the *load factor*. From the standpoint of economics as well as logistics, a relatively flat load duration curve with a high load factor is clearly desirable for utilities. This is because the cost of providing service consists in large part of investments related to peak capacity, whereas revenues are generally related to total energy consumed (i.e., average demand). A pronounced peak indicates a considerable effort that the service provider must undertake to meet demand on just a few occasions, although the assets required to accomplish this will tend not to be utilized much during the remainder of the year. For example, in the case illustrated in Figure 6.4, note that the resources required to meet the top 5000 MW of demand—roughly 10% of the total capacity investment—are called upon for less than 1% of the year, or all of three days.

The load factor obviously depends on climate, but it also depends on the diversity within the customer base, or *load diversity*. For example, commercial loads that operate during the day may be complemented by residential loads before and after work hours. Improvement of the load factor through increased load diversity was a major factor in the historical expansion of power systems, as was the ability to share resources for meeting the peak.

More recently, an important concern about daily load profiles is the *ramp rate*, or rate of change (say, in megawatts per hour). This issue is especially salient in the presence of variable generation from renewable resources such as solar and wind, which are not dispatchable by the system operator or utility. Solar generation, which now contributes a substantial fraction of power and energy supply in many areas, can create a distinct mid-day reduction of demand, as shown in Figure 6.5, along with a steep increase around sunset (creating what has been called the “duck curve” for its visual appearance).

System demand minus wind and solar, in 5-minute increments, compared to total system and forecasted demand.

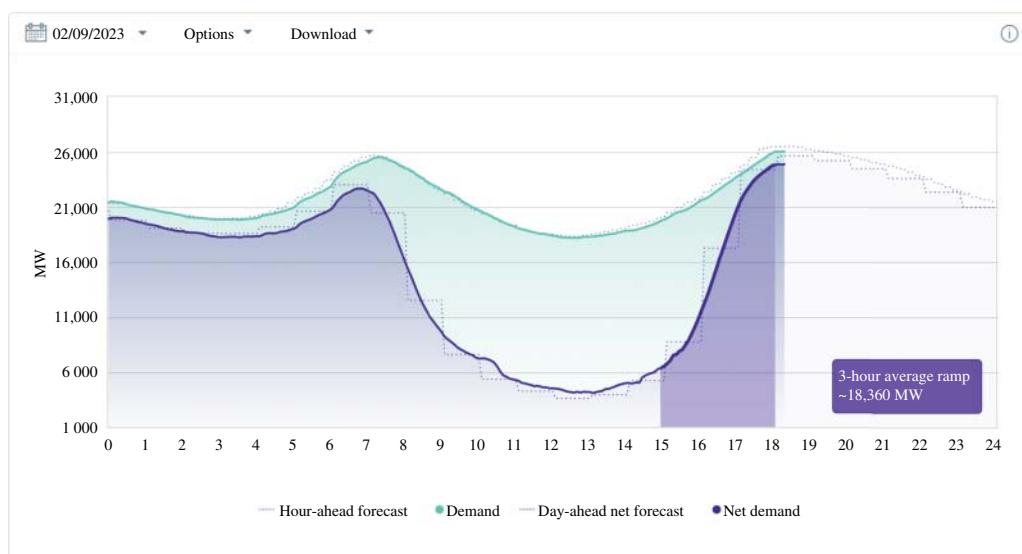


Figure 6.5 Load profile on a mild, sunny day. Source: Adapted from California Independent System Operator.

Net demand shows customer load minus nondispatchable generation that is monitored and accounted for on the supply side. Demand by itself reflects what is measured at customer meters, which may include *behind-the-meter* generation from rooftop PV that is not separately measurable by the utility. In the case of California, the demand profile on a cool, sunny day in early spring clearly shows the effect of daytime solar generation as *negative load*. The resulting steep increase in supply needed to make up for diminishing sunshine in the afternoon introduces new operational and economic challenges, especially since generators to help out may not be highly utilized at other times and thus sustain a poor *capacity factor* (the mirror image of load factor, on the generation side). Other approaches to mitigate ramp rates include energy storage (see Section 15.3) and more active measures to manage load than have been used historically.

6.4.4 Managing Load

A range of strategies and methods can be applied to control electric loads directly, or to create incentives for customers to do so, in a way that is responsive to the needs and constraints of the electric grid. Such strategies are sometimes called *demand response*, or *demand-side management* from the utility standpoint.

First, it is important to distinguish demand response from *energy efficiency* and *conservation*. The term “conservation,” popularized in the United States during the energy crisis of the 1970s, has fallen out of use because of its implication of frugality, limitation, and associated discomfort: to conserve energy means to use less of it, presumably at some inconvenience (such as sitting in a cold room with a sweater). The term “efficiency” implies instead that the same convenience or comfort level is enjoyed with less energy consumption (like attaining the same temperature in the house with a better heating system and better wall insulation). Thus, efficiency is about reducing waste, not service.

Some important examples of loads with greatly improved efficiency as of the 2020s include LED lighting, variable-speed motor drives, and heat pumps for water and space heating. All of these use different physical principles of operation than their predecessors (incandescent lamps, induction motors, and resistive heating), but the results delivered to meet the customer’s end use needs are practically indistinguishable. Programs to accelerate customer adoption rates of energy efficient technologies have been one important strategic resource for governments and utilities in the effort to reconcile electric supply and demand at minimum overall cost to society, while sustaining economic growth.¹⁷

By contrast, demand response is not concerned with reducing electric energy consumption in general, but rather with manipulating the temporal profile of electric power demand. The effects of efficiency and demand response overlap in that they both serve to reduce peak demand, but the expectation in the latter case is that the demand is merely delayed, so that the load will rebound at a later time when the grid can better accommodate it.

The time frame for demand response can range from seconds to days, but most typically minutes to hours. A report¹⁸ led by the Lawrence Berkeley National Laboratory introduced the memorably alliterative terminology of “shape, shift, shed and shimmy.” Load *shaping* refers to consistent changes in consumption patterns that alter the daily demand profile—for example, through

17 The problem of designing the most effective incentives and mechanisms to apply future energy savings to mitigate the higher upfront cost of many efficient appliances is a rich subject for energy policy studies.

18 Peter Alstone, Jennifer Potter, Mary Ann Piette, Peter Schwartz, Michael A. Berger, Laurel N. Dunn, Sarah J. Smith, Michael D. Sohn, Arian Aghajanzadeh, Sofia Stensson, Julia Szinai, Travis Walter, Lucy McKenzie, Luke Lavin, Brendan Schneiderman, et al., 2025 California Demand Response Potential Study—Charting California’s Demand Response Future. LBNL-2001113 (2017).

time-of-use rates that generally discourage late afternoon demand—with plenty of advance notice. Load *shifting* describes changes in response to real-time alerts or specific price signals to address acute grid constraints, where load is moved to a more desirable time of day. To *shed* load means for the utility or grid operator to forcibly disconnect or *curtail* customers, whose tariff may have accounted for that possibility. Finally, the term *shimmy* (which alludes to a dance move) refers to fast, dynamic load response that may be credited as *ancillary services* such as load following and frequency regulation (see Section 11.1).

Some loads naturally lend themselves to being shifted or shimmied without noticeable impact on the end user, because of their inherent time delay or storage characteristics. These include battery charging, and heating or cooling applications. Electric vehicle charging is a prime candidate for aggressive load management for several reasons: the sheer size of the load (where one car may be equivalent to the entire rest of a household's electricity use), its concentration (which may also impact local voltage or power quality on the grid, so that active control offers multiple benefits), and the fact that EV chargers contain programmable power electronics that can readily be adapted to respond to grid needs as the technology proliferates. The time scale and depth of EV load shifting or shimmying can vary depending on user flexibility and depending on trade-offs against battery degradation. In the *vehicle-to-grid* (V2G) option, EV batteries are doubly recruited as a controllable load and storage asset that can inject power to the grid on demand.

Another class of loads that lend themselves to active management are *thermostatically controlled loads* (TCLs), as noted under *Thermal Storage* in Section 15.3.3. Examples include space and water heating, air conditioning, commercial and industrial chilling, and some types of industrial process heating where time is not of the essence. The idea is to take advantage of the thermal inertia within the physical medium to be heated or cooled, so that the instantaneous electric power demand can be responsive to the needs of the grid while the target temperature stays within acceptable bounds. With clever coordination, large numbers of small TCLs could be aggregated to offer a significant resource that can regulate demand up as well as down.

Fast and flexible demand response in particular hinges on enabling technologies. The first crucial step toward load shaping was time-differentiated metering, also known as “smart meters,” that record instantaneous demand in kilowatts rather than just cumulative energy consumption in kilowatt-hours. These meters were a crucial prerequisite for *time-of-use* (TOU) rates that discriminate among predetermined on- and off-peak periods.

More sophisticated tools include communications for real-time price or control signals at the level of a customer, building, or individual appliance. At higher levels of granularity, software becomes an important component of a demand response infrastructure, as more optimization and control is delegated to automated algorithms to shield customers from having to pay constant attention. For example, a home energy management system may communicate with individual devices while presenting a single interface to the user or the utility.

Interestingly, advances in information technology and communication infrastructure also open new possibilities for diverse entities to participate in the orchestration of resources. For example, third-party aggregators may support the logistics for many small households to collectively bid demand response into a power market. These opportunities also come with new challenges such as privacy, trust, and cyber-security.

Such modern approaches stand in contrast to *interruptible tariffs*, an early and crude form of demand response where the utility simply and directly shuts off service to participating individual customers at the meter during grid emergencies. Interruptions may be limited by contract to some

number of events per year, but the customer has no particular control over these occurrences.¹⁹ Interruptible tariffs have been available mainly to larger customers willing to trade service continuity for substantially reduced rates. Refinements enabled by more granular communication and control imply the possibility of both spreading the response across larger numbers of customers, and affording them a more active role in deciding what loads to turn off when.²⁰

Controllable load can be treated either as a form of supply to the grid, or as demand through sophisticated tariffs that offer detailed incentives for time-varying behavior.²¹ One difference is that from a supply perspective, it is necessary to establish a *baseline* against which a purposeful reduction in power or energy demand is measured. In any case, given the technical potential and often favorable costs compared to increasing supply, demand-side resources seem destined to play a much larger role in grid operations than they have historically.

Problems and Questions

- 6.1** Consider a simple, old-fashioned rheostat connected in series with an incandescent light bulb. The supply voltage to the series combination remains constant. Find the setting, expressed as a fraction of the light bulb's resistance, at which the rheostat itself dissipates the most heat.
- 6.2** Predict the effect on (a) real power demand, (b) reactive power demand, and (c) energy consumption by a collection of purely resistive electric water heaters, if service voltage is reduced by 2.5% (to 97.5% of nominal).
- 6.3** Suppose that the water heaters in the previous problem are replaced by heat pump water heaters (i.e., compressor motors). Qualitatively, how would you expect your answers to change?
- 6.4** An electronic ballast for a fluorescent light is characterized by the following ZIP coefficients: $Z_p = 0.22$, $I_p = -0.5$, $P_p = 1.28$, $Z_q = 9.64$, $I_q = -21.59$, $P_q = 12.95$. By how much do real and reactive power demand for this load change if the voltage (a) drops to 90% of nominal, or (b) increases to 110%?
- 6.5** A load duration curve is given by the following information: The peak load of 10 MW occurs 20% of the time. 60% of the time, the load is 8 MW, and the remaining 20% of the time, the load is 1 MW.
 - (a) Draw a rough sketch of the LDC.
 - (b) What is the load factor?

¹⁹ The idea is analogous to airline passengers accepting payment in return for yielding their seats on an overbooked flight, except that the electric customer faces only the possibility, not the certainty of interrupted service. Indeed, anecdotal evidence suggests that participating customers are sometimes surprised and dismayed when an interruption actually occurs.

²⁰ For example, “smart panels” can physically limit power demand at the meter to some number of kilowatts, while the customer selects which circuits to prioritize on a smartphone app.

²¹ Intuitively, and also from a theoretical standpoint in economics, it seems that markets should treat supply and demand symmetrically: if the objective is simply to balance generation and load, we might be indifferent as to whether this is accomplished by adjusting generation or adjusting load, and an increment or decrement of power might be valued equally from either resource. In practice, it is not so simple, and the details are well beyond the scope of this text.

- 6.6** In the previous scenario, a new load of 4 MW is added. It operates only 1% of the time, but always coincident with the previous 10-MW peak.
- What is the new load factor, to two significant figures?
 - By what percentage has energy consumption grown?
 - In your own words, explain why this load addition could be costly for the electricity supplier.

7

Transmission and Distribution Systems

7.1 System Structure

7.1.1 Interconnection

Since the beginnings of commercial electric power in the 1880s, the systems for its delivery from production sites to end users have become increasingly large and interconnected. Although the 21st century is also seeing small-scale microgrids as a complementary development, the interconnection trend and the reasons for it remain relevant. In the early days, the standard “power system” consisted of an individual generator connected to an appropriately matched load, such as Edison’s famous Pearl Street Station in New York City that served a number of factories, residences, and street lighting. The trend since the early 1900s has been to interconnect these isolated systems with each other, in addition to expanding them geographically to capture an increasing number of customers. Owing to a considerable investment in Public Works projects for rural electrification, most U.S. citizens had electricity by World War II. The process of interconnecting regional systems into an expansive synchronous grid has continued throughout the postwar era, leaving us today with only three electrically separate alternating current (a.c.) systems in the United States: the Western United States, the Eastern United States, and Texas.¹ Similarly, the Western European system is completely interconnected and synchronous from Portugal to Denmark, Austria, and Italy.

The continuing geographical expansion and interconnection of power systems over the course of the last century has been motivated by a variety of technical, social, and economic factors. For example, early drivers included a sense of cultural progress associated with a connected grid. In recent decades, economies of exchange, or opportunities for sales of electricity, have been a key motivator for strengthening transmission interconnections or *interties* between regions. The main reasons for expansion and interconnection have been threefold: *economies of scale*, improvement of the *load factor*, and enhancement of reliability by pooling generating *reserves*. Another contemporary motivation for further strengthening long-distance grid interconnection is to access the most abundant solar and wind resources.

An economy of scale simply means that it tends to be less expensive to build and operate one large unit than several smaller ones. This makes sense because much of the construction process of conventional power plants—from design and licensing to pouring concrete and bringing in the crane that lifts the generator—tends to involve fixed rather than variable costs, not depending very much on the unit’s megawatt capacity. Even operating costs, though dominated by fuel, have economies of scale in aspects such as labor, maintenance, and operational support. Figure 7.1

¹ The state of Texas opted to forgo connectedness in the interest of avoiding interstate commerce regulations and federal oversight.

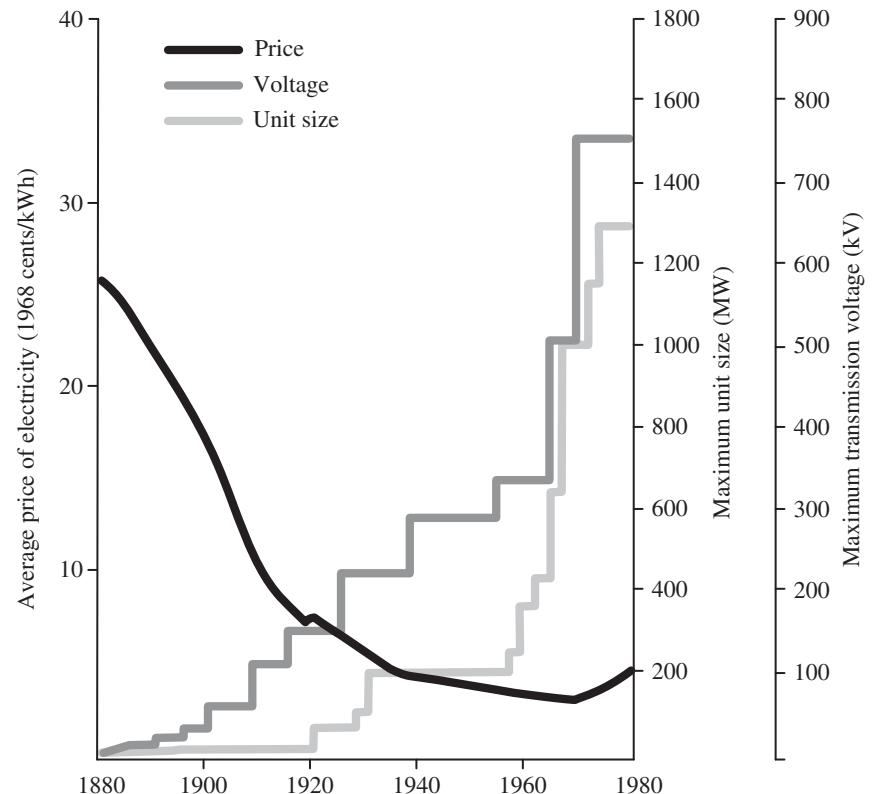


Figure 7.1 Historical growth of generation unit size and transmission voltage. Source: Adapted from Economic Regulatory Administration, 1981.

shows the historical increase in the size of generation units as a result. The graph also reflects the limit that appears to have been reached in the late 1960s in terms of the maximum practicable and efficient unit size, marking the end of an era of declining costs.² In the early days of power systems, however, the advantage of increased unit size was substantial and provided an important incentive for utilities to connect enough customers so as to take full advantage of economies of scale.³

The load factor is the ratio of a load's actual energy consumption over a period of time to the maximum amount of power it demands at any one instant (also see Section 6.4). This is a key criterion for the economic viability of providing electric service, since the cost of building the supply infrastructure is related to the maximum amount of power (i.e., the capacity of generators and transmission lines), whereas the revenues from electricity sales are related to the amount of energy (kilowatt-hours) consumed. Thus, from the supply standpoint, the ideal customer would be demanding a constant amount of power 24 hours a day. Of course, this does not match the actual

2 This point is carefully documented by Richard Hirsh, *Technology and Transformation in the American Electric Utility Industry* (Cambridge, England: Cambridge University Press, 1989). Hirsh argues that unexpectedly diminishing economies of scale played an important role in precipitating the economic crisis that U.S. utilities experienced in the 1970s, independent of the widely recognized factors of increasing fuel costs and unprofitable investments in nuclear power.

3 Modern technologies, especially solar photovoltaics (PV), are highly modular and can be used over a wide range of sizes to capture locational benefits (see Section 15.2 on Distributed Generation). Still, economies of scale favor large plants for bulk energy.

usage profile of real customers; nevertheless, a smoother consumption profile can be accomplished by *aggregating* loads, that is, combining a larger number and different types of customers within the same supply system whose times of power demand do not coincide.

For example, an individual refrigerator cycles on and off, using a certain amount of power during the time interval when it is on, and none the rest of the time. But if a number of refrigerators are considered together, their cycles will not all coincide; rather, they will tend to be randomly distributed over time. The larger the number of individual loads thus combined, the greater the driving force of statistics to level out the sum of power demand. By expanding their customer base to include both a larger number of customers and customers with different types of needs (such as machinery that operates during business hours versus lighting that is needed at night) so as to deliberately combine complementary loads, utilities improved their load factor and increased their revenues in relation to the infrastructure investment. Accordingly, electric power systems grew from the scale of city neighborhoods to cover entire counties and states.⁴

The third factor driving geographical expansion and interconnection has to do with the ability to provide greater service reliability in relation to cost. The basic idea is that when a generator is unavailable for whatever reason, the load can be served from another generator elsewhere. To allow for unexpected losses of generation power or *outages*, utilities or balancing authorities maintain a *reserve margin* of generation, standing by in case of need. Considering a larger combined service area of several utilities, though, the probability of their reserves being needed simultaneously is comparatively small. If neighboring utilities interconnect their transmission systems in a way that enables them to draw on each other's generation reserves, they can effectively share their reserves, each requiring a smaller percentage reserve margin at a given level of reliability.

More extensive interconnection of power systems also provides for more options in choosing the least expensive generators to dispatch, or, conversely, for those with a surplus of inexpensive generating capacity to sell their electricity. For example, the north-south interconnection along the west coast of the United States allows the import of hydropower from the Columbia River system down through California. In general, over the course of the development of electric power systems during the 20th century, larger and more interconnected systems have expanded the options for managing and utilizing resources for electric supply in the most economic way.

As the distance spanned by transmission lines increased, so did the importance of energy losses due to resistive heating. Recall from Section 1.4.2 that high voltage is desirable for power transmission in order to reduce current flow and therefore resistive losses in the lines. Therefore, the geographical expansion of electric grids gave an increasing incentive to operate transmission lines at higher voltages. The maximum voltages used for transmission steadily increased over the decades, as shown in Figure 7.1 (note that the first scale on the left indicates unit size in megawatts; the scale on the far right transmission voltage in kilovolts).

There are also liabilities associated with larger size and interconnection of power systems. Long a.c. transmission lines introduce the problem of stability (Chapter 13). More interdependence between and among areas also means greater vulnerability to disturbances far away, including voltage and frequency fluctuations. Conventional wisdom in the electric power industry holds that the benefits of interconnection outweigh the drawbacks, at least up to the scale at which power systems are presently operated. Enabled by significant recent technological innovation, *microgrids* (Section 15.4) offer a complementary approach for supporting reliability on a very local basis.

⁴ This argument is developed in detail with ample historical illustration by Thomas P. Hughes, *Networks of Power: Electrification in Western Society, 1880–1930* (Baltimore, MD: Johns Hopkins University Press, 1983).

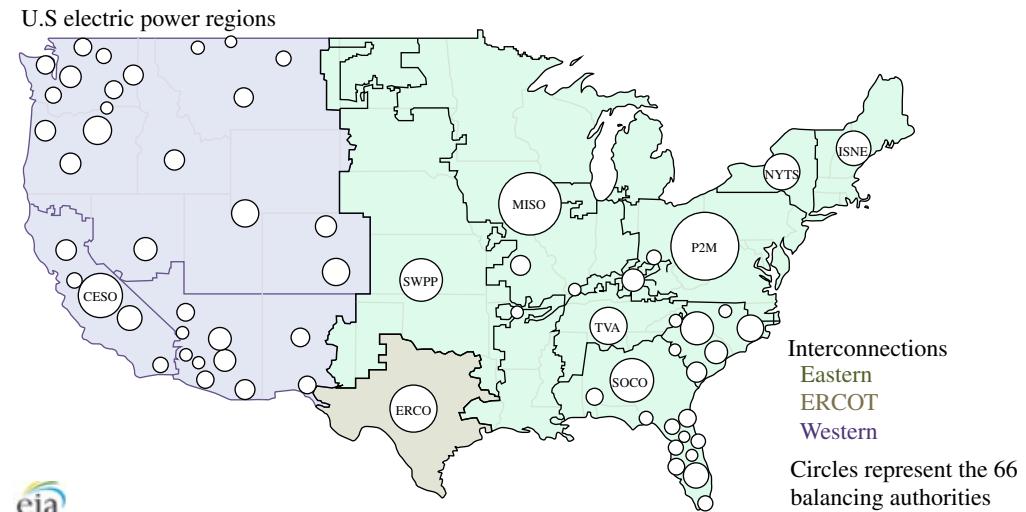


Figure 7.2 Regions and interconnections of the U.S. electric grid. Source: U.S. Energy Information Agency, 2016/Public Domain.

Overall, while individual utilities retain their geographically delimited service territories, the strength and importance of regional interties and power exchange has steadily increased over the past decades.⁵ Meanwhile, restructuring of the vertically integrated industry (Section 16.4.2) has assigned different responsibilities to different entities, with many utilities focused primarily on distribution system services. From the standpoint of transmission-level operations and coordinating power exchange over long distances, relevant organizational entities include *Independent System Operators* (ISOs), *Balancing Authorities* (Section 11.1.5), *Regional Transmission Operators* (RTOs), and regional *reliability councils*. Because the success of system operation depends on close technical cooperation, such groups are administratively organized at various levels, under private or public agencies. For the United States and Canada, the not-for-profit North American Electric Reliability Corporation (NERC) serves as the overarching Electric Reliability Organization. RTOs and most ISOs in the United States⁶ are regulated by the Federal Energy Regulatory Commission (FERC). Figure 7.2 maps some of these entities.

Notwithstanding the overall integration trend, the U.S. grid remains divided into three major interconnections: the Eastern Interconnection, the Western Electricity Coordinating Council (WECC), and the Electric Reliability Council of Texas (ERCOT). Each interconnection is internally synchronous, meaning that all generators are physically coupled by a common a.c. frequency and jointly performing *load frequency control* (Section 11.1). The three interconnections are coupled to each other only by direct-current links of comparatively small transfer capacity.

Figure 7.3 illustrates the separation between the interconnections in terms of their distinct a.c. frequencies.⁷

5 This statement applies to the United States as well as other countries, notably within the European Union.

6 With the exception of ERCOT, regulated by the Public Utilities Commission of Texas.

7 This snapshot of a real-time frequency map also reveals that “steady-state frequency” is in fact an approximation, as dynamics result in slightly different frequencies reported by phasor measurement units (Section 16.2.4) at different locations within a synchronous area. The answer to the question, “True or false: The a.c. frequency is the same everywhere within a synchronous grid?” is, “It depends on how closely you’re looking.”

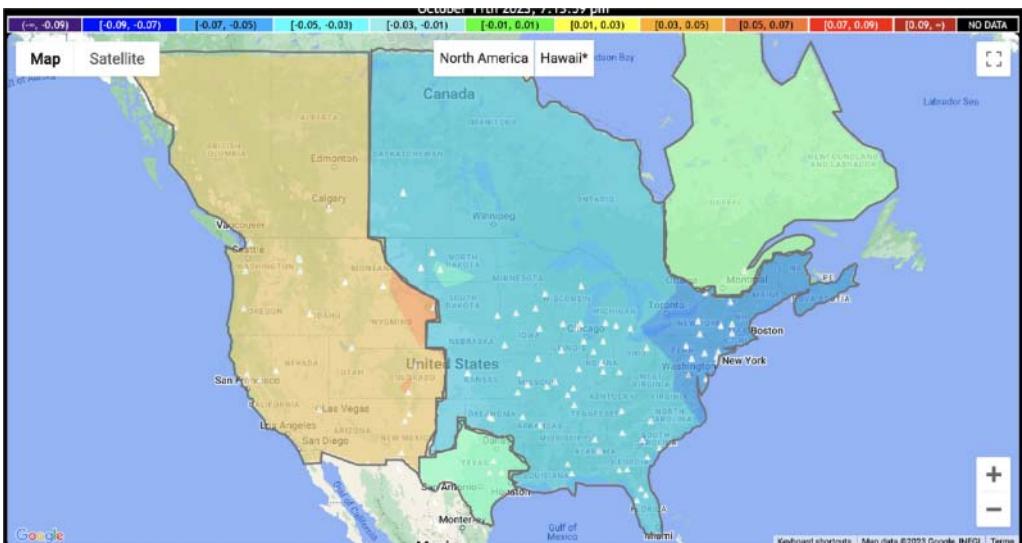


Figure 7.3 Real-time frequency in the major synchronous interconnections in North America, as seen on FNET (University of Tennessee, Knoxville and Oak Ridge National Laboratory). The variations indicated on the color legend represent fractions of a hertz above or below the nominal 60.00 Hz. Source: The University of Tennessee Knoxville/<https://fnetpublic.utk.edu/frequencymap>.

In view of a decarbonized grid (Section 16.4.1), exploiting the most abundant wind resources in the central plains and solar resources in the southern United States will require strengthening transmission links to population and load centers. Due to the long distances involved as well as the need to cross between interconnections that are asynchronous with each other, high-voltage d.c. transmission (Section 7.2.4) is an important technology in this context.

7.1.2 Structural Features

High voltages are crucial for power transmission over long distances. Closer to the end users of electricity, however, safety prohibits the use of equipment at excessively high voltages, lest people start electrical fires or electrocute themselves (a point that every parent of a curious toddler can especially appreciate). The sheer expense of insulating and ensuring proper clearance around high-voltage equipment is also a consideration. In the design of power delivery systems, the greater energy efficiency of high voltage and low current must therefore be weighed against safety and capital cost. Rather than having to settle for some intermediate voltage as a compromise, the use of *transformers* (Chapter 8) makes it possible to operate different parts of the system at different voltages, retaining the respective advantages of higher and lower voltages in those places where they are most consequential.

Power delivery systems are therefore divided into two general tiers: a *transmission system* that spans long distances at high voltages on the order of hundreds of kilovolts (kV), usually between 60 and 500 kV, and a more local *distribution system* at intermediate voltages in the low tens of kV. The latter is more specifically referred to as *primary distribution*, in contrast to the *secondary distribution system*, which consists of the wires that directly connect most domestic and small commercial customers, at voltages in the 100-V range (nominally 120 V in the United States). Larger commercial and industrial customers often receive their service at higher voltages, connected directly into the primary distribution system. The transmission system is also further

subdivided into *subtransmission*, operated in the neighborhood of 100 kV, and longer-distance transmission at several hundred kV.⁸ Collectively, the entire power delivery system is referred to as the *transmission and distribution (T&D) system*.

The division between transmission and distribution is defined in terms of voltage level, though individual utilities have different customs for where, exactly, to draw the line. In general, “distribution” means below 60 or 70 kV. Physically, the boundary between transmission and distribution systems is demarcated by transformers, grouped at *distribution substations* along with other equipment such as circuit breakers and monitoring instrumentation. Organizationally, most utility companies have separate corporate divisions responsible for the operation and maintenance of transmission versus distribution systems; the substations themselves may be considered to lie on either side, but most often fall under the jurisdiction of power distribution.

Figure 7.4 shows the basic structure and components of a transmission and distribution system. First, note that the illustration is not drawn to scale. Note also that it is a *one-line diagram*, which does not show the three phases for each circuit (Chapter 4). The vertical lines represent *buses*, or common connections at key points in the system, especially power plants and substations. With power flowing from left to right, the diagram indicates the hierarchical relationship among the important subsystems, along with some typical voltage values.

On the far left side of the diagram, two generators deliver power at 21 kV. They connect to the transmission system through a transformer (indicated by the wavy symbol reminiscent of wire coils) that steps up the voltage to 230 kV. The squares on either side of the step-up transformer indicate circuit breakers that can be opened to isolate the generator from the system. Circuit breakers also appear elsewhere in the diagram; their function is discussed in Section 7.5 on protection.

The system shown includes both high-voltage transmission at 230 kV and subtransmission at 60 kV. The transmission and subtransmission systems meet in a transformer at a transmission substation. At the distribution substation, the voltage is stepped down further to the primary distribution voltage, in this case 12 kV. The primary distribution lines or *feeders* branch out from the substation to serve local areas. These main feeders carry all three phases (see Section 4.1).

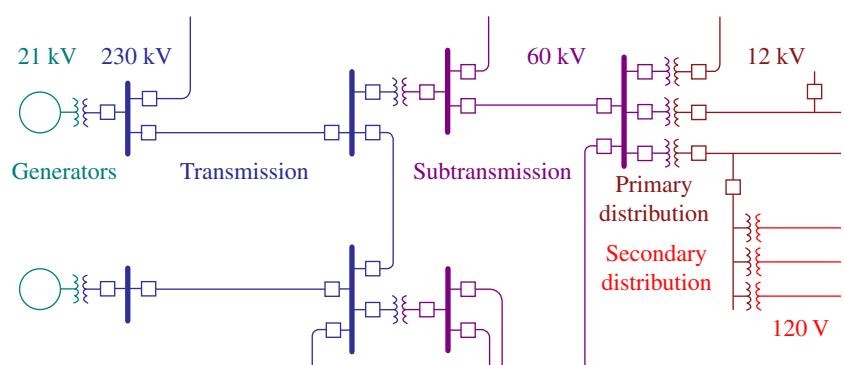


Figure 7.4 One-line diagram showing basic power system structure.

⁸ In European usage, the terms translate as follows: subtransmission voltage = high voltage; primary voltage = medium voltage; and secondary voltage = low voltage. Owing to the smaller distances in Europe (where line losses play a less important role), “high voltage” is generally limited to 380 kV, whereas 500 and 750 kV are used in the United States. Worldwide, the highest voltages used to date are 1200 kV a.c. for some very long transmission lines across Siberia and ± 600 kV d.c. in Brazil.

From the main feeders, *lateral* feeders (*laterals* for short) carry one or two phases for a shorter distance—a few city blocks, for example. From the lateral, several *distribution transformers* step the voltage down again to the secondary level at which most customers are served, generally 120 V. Typically, one distribution transformer serves several residences, up to one city block. Larger commercial or industrial consumers (not shown here) usually have three-phase service directly from the primary distribution level through their own dedicated transformer.

Transmission and distribution infrastructure may be overhead (often abbreviated OH) or underground (UG). Underground lines are substantially more expensive, roughly by an order of magnitude (factor of 10). Contributing to the high cost are not only trenching, but the need for insulated cable and the fact that without convective cooling, the allowable current rating for a given conductor size underground is much less. High-voltage transmission lines are very rarely underground except for short distances in high-density urban areas or near airports. A different emerging motivation to consider the expense of undergrounding transmission lines is to avoid fire hazards in extremely fire-prone areas, where the smallest spark (which might occur due to vegetation contact, or some mechanical problem with the line) can spell disaster.

In the United States, distribution systems are often underground in high-density urban areas, in housing developments where entirely new roads and infrastructure are built, in affluent neighborhoods that might share some of the cost, and in areas with high fire risk. In these situations, transformers and switchgear are also typically underground in vaults, adding to the cost. Underground distribution systems tend to experience significantly fewer outages since they are not as exposed and vulnerable to the elements, vegetation, and animals. However, if there is a problem with the underground infrastructure, it can be more difficult to locate.

7.1.3 International Differences in Distribution System Design

In Europe, as well as in many countries formerly colonized by Europeans, power distribution systems have a distinctively different look. Unlike the United States, where there is usually a distribution transformer (typically single phase) for every few customers, connected by short service drops, in Europe there are fewer and larger transformers—usually hidden in vaults rather than mounted on poles—from which a more extensive system of secondary lines branches out. Because secondary (low-voltage) lines are less expensive to underground than the primary (medium-voltage) lines that make up the larger part of the U.S.-style distribution system, distribution systems are more often undergrounded and thus altogether less visible in Europe. These differences in design are consistent with differences in geography, population and load density, and the historical expansion of power systems.

Except when compared with downtown urban areas, the population density in Europe is generally higher and individual loads smaller, making it more feasible to extend secondary lines to many customers. Extension of secondary lines is also more feasible because the standard secondary voltage is higher in Europe (nominally 220 V). This design approach makes it relatively easy to geographically extend service areas by adding another secondary line to an existing transformer, though it becomes more difficult once the transformer capacity is insufficient. The layout of the system is generally optimized around load level.

In North America, particularly in rural areas, the location of customers and the size of the area to be covered are generally more important considerations in determining system layout than load level. Because customers tend to be much farther apart, higher-voltage (primary) lines are needed to reach them. Systems here tend to have a higher loading capacity per mile of circuit, and load growth within the existing service area can usually be quite readily accommodated by adding

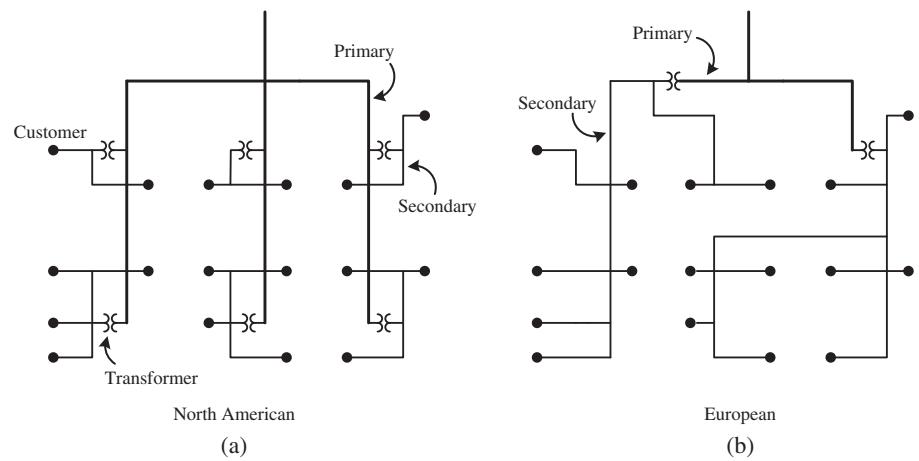


Figure 7.5 North American (a) and European (b) distribution systems.

small transformers.⁹ Figure 7.5 illustrates the two different types of layout (adapted from Carr and McCall, 1992), where the heavier lines graphically represent primary and the lighter lines secondary distribution wires. Note that the line weight does not reflect physical wire diameter, since a lower-voltage line actually carries a greater current to transmit a given amount of power.

7.1.4 Stations and Substations

Transmission and distribution stations exist at various scales throughout a power system. In general, they represent an interface between different levels or sections of the power system, with the capability to switch or reconfigure the connections among various transmission and distribution lines. On the largest scale, a transmission substation would be the meeting place for different high-voltage transmission circuits. At the intermediate scale, a large distribution station would receive high-voltage transmission on one side and provide power to a set of primary distribution circuits. Depending on the territory, the number of circuits may vary from just a few to a dozen or so. The major stations include a control room from which operations are coordinated. Smaller distribution substations follow the same principle of receiving power at higher voltage on one side and sending out a number of distribution feeders at lower voltage on the other, but they serve a more limited local area and are generally unstaffed.

The central component of the substation is the *transformer*, as it provides the effective interface between the high- and low-voltage parts of the system. Other crucial components are *circuit breakers* and *switches*. Breakers serve as protective devices that open automatically in the event of a fault, that is, when a protective relay indicates excessive current due to some abnormal condition. Switches are control devices that can be opened or closed deliberately to establish or break a connection. An important difference between circuit breakers and switches is that breakers are designed to interrupt abnormally high currents (as they occur only in those very situations for which circuit protection is needed), whereas regular switches are designed to be operable under normal currents. Breakers are placed on both the high- and low-voltage sides of transformers. Finally, substations may also include capacitor banks to provide voltage support. Because three phases are used, all equipment comes in sets of three, referred to as a “bank” for example, a transformer bank is a set of three transformers, one for each phase. (A fourth may be kept adjacent, as a spare).

⁹ J. Carr and L.V. McCall, “Divergent Evolution and Resulting Characteristics among the World’s Distribution Systems”, *IEEE Transactions on Power Delivery* 7(3), 1601–1609, July 1992.

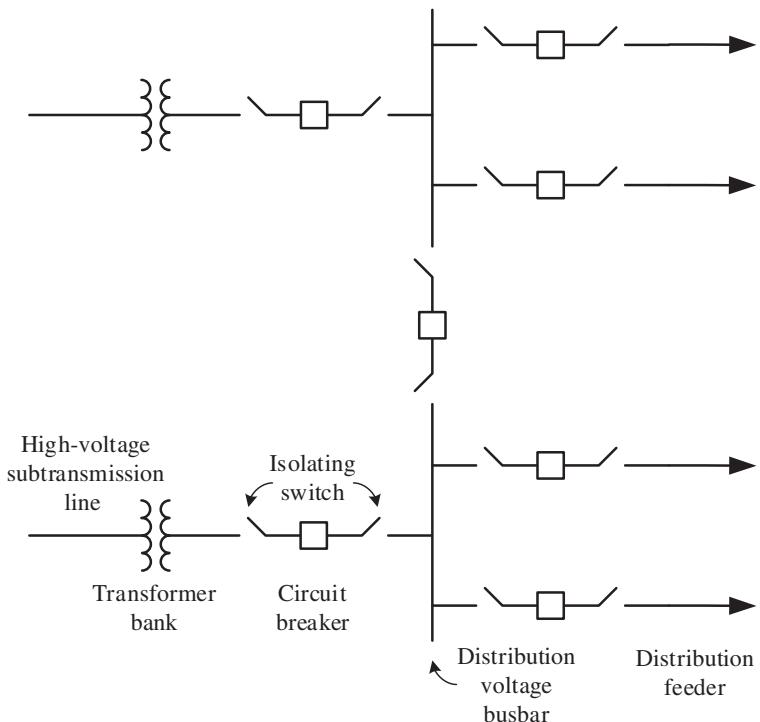


Figure 7.6 Distribution substation layout.

A sample layout for a distribution substation is shown in Figure 7.6. Note that schematic diagrams like this are not drawn to scale, nor do they necessarily provide a good sense of the actual spatial arrangement of the switchyard for those unfamiliar with the physical equipment represented here. Figure 7.7 shows a photograph of a similar substation.

7.1.5 Topology

An important characteristic of transmission and distribution systems is their topology, or how their lines are connected. The most important distinction is between a *radial* configuration where lines branch out sequentially and power flows strictly in one direction, and a *network* configuration that is more interconnected. In a network, any two points are usually connected by more than one path, meaning that some lines form loops within the system.

Transmission systems are generally networks. Local portions of a transmission system can be radial in structure—for example, the simplified section shown in Figure 7.4, with all the power being fed from only one side. Since generating plants are more likely to be scattered about the service territory, though, the system must be designed so that power can be injected at various locations and power can flow in different directions along the major transmission lines, as necessitated by area loads and plant availability. Thus, high-voltage transmission systems consist of interconnected lines without a hierarchy that would distinguish a “front” or “back” end.

It is true, of course, that due to the geography of generation and major load centers, power will often tend to flow in one direction and not the other. Nevertheless, this kind of directionality is not built into the transmission hardware. For example, in the state of New York, power tends to flow from north to south into New York City, but as far as the transmission system is concerned, the power could just as easily be sent from south to north. The system structure does finally become



Figure 7.7 Distribution substation with transformers of different vintage, supplied from 60 kV subtransmission (left). Also visible are voltage regulators (back right), small service transformers (far right), and simple mechanical switches (operated with a hookstick).

hierarchical at the interfaces with the lower-voltage subsystems (subtransmission or distribution), where power is intended to flow only from high to low voltage.

The network character of the transmission system allows for different operating conditions in which power may flow in different directions. It also offers the crucial advantage of *redundancy*. Because there are multiple paths for power to flow, if one transmission line is lost for any reason, all the load can still be served (as long as the remaining lines can carry the additional load). Indeed, a standard design and operating criterion for transmission systems requires that the network as a whole must continue to function if any one link is interrupted.

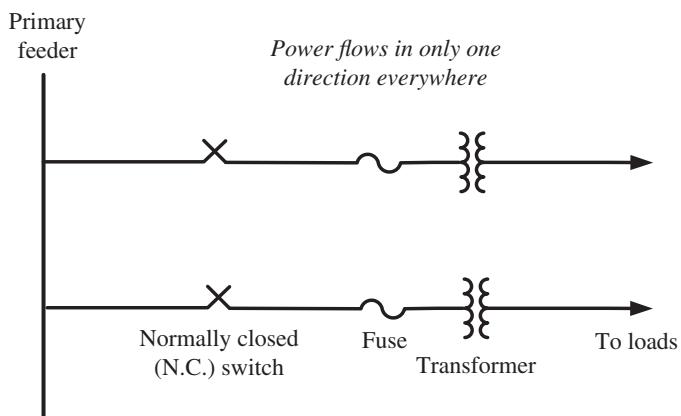


Figure 7.8 Radial distribution system.

The basic radial design concept is illustrated in Figure 7.8. The radial system has a strict hierarchy: power flows only in one direction; there is always an “upstream” and a “downstream.” The distribution lines or feeders extend and branch out in all directions from a substation somewhat like spokes from a hub. Owing to this hierarchy, any given line or component can only be energized from one direction. This property is crucial in the context of circuit *protection*, which means the interruption of circuits or isolation of sections in the event of a problem or *fault*. In a radial system, circuit breakers can readily be located so as to isolate a fault—for example, a downed line—immediately upstream of the problem, interrupting service to all downstream components. Economically, radial systems also have the advantage that smaller conductor sizes can be used toward the ends of the feeders, as the remaining load connected downstream diminishes.

Figure 7.9 illustrates a *loop system*. One switch near the midpoint of the loop is open (labeled N.O. for normally open) and effectively separates the loop into two radial feeders, one fed by each transformer. Thus, under normal operation, the sections are not connected at that point, so that the system operates as a radial system. But under certain conditions—for example, a failure of one of the two substation transformers—the N.O. switch can be closed and one section of the distribution system energized through the other. By choosing which one of the other switches to open, sections of the loop can be alternatively energized from the left or right side. This has the advantage of enabling one transformer to pick up additional load if the other is overloaded or out of service, and of restoring service to customers on both sides of a fault somewhere on the loop. While loops are operated as radial systems at any given time, that is, with power flowing only outward from the substation transformer, the hardware including protective devices must be designed for power flow in either direction.

Finally, Figure 7.10 shows a simple example of a network distribution system. A networked system is generally more reliable because of the built-in redundancy: if one line or transformer fails, there is another path for the power to flow. The cost of a network system to serve a given area is higher than a simple radial system, owing not only to the number of lines but also the necessary equipment for switching and protection. Networks are often used in downtown metropolitan

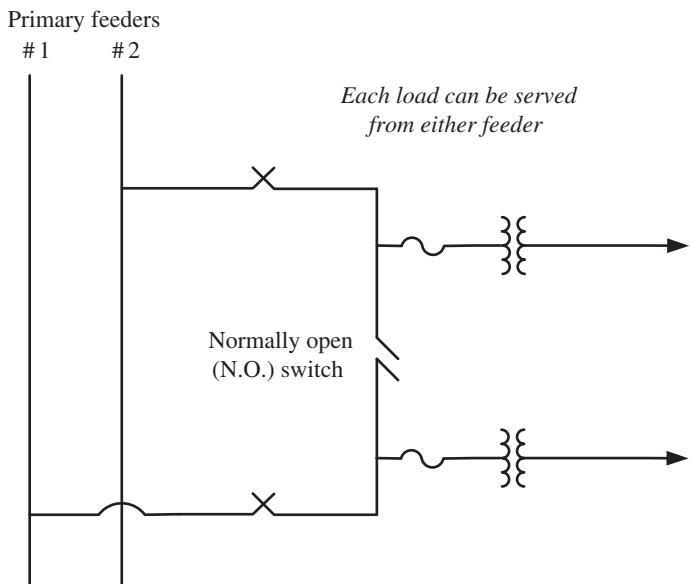
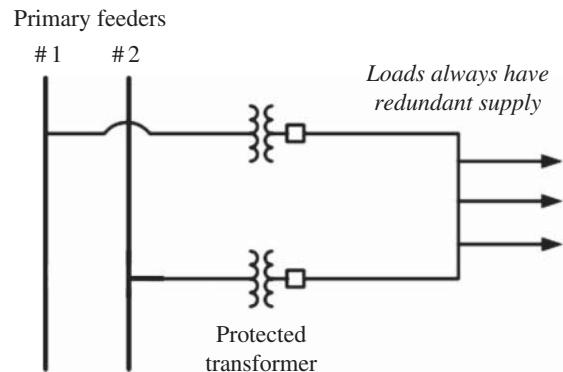


Figure 7.9 Loop system.

**Figure 7.10** Networked topology.

areas where reliability is considered extremely important, and where the load density justifies the capital expense.

From the standpoint of circuit protection, a network is much more challenging because there is no intrinsic upstream or downstream direction, meaning that a given point in the system could be energized or receiving power from either side. This means that any problem must be isolated on *both* sides, rather than just on the upstream side. However, the objective is still to make the separation as close to the fault as possible so as to minimize the number of customers affected by service interruptions. As a result, the problem of coordinating the operation of multiple circuit breakers becomes more complicated (Section 7.5.2).

Even under normal operation, a network requires preventive measures against circulating currents or *loop flows* (see Section 7.1.7). For example, if the voltage on the primary feeders #1 and #2 is even slightly different, a current will flow from one to the other through the two transformers, unless they are specifically protected against reverse power flow. Since that circulating current would be limited only by the impedance of the local network, which generally should be small, the current could become quite large, and even if it is not immediately dangerous it entails losses. Protected network transformers with directional relays will block the reverse current, but they are significantly more expensive.¹⁰ For these reasons, most distribution systems either have a strictly radial layout or operate in radial configuration at any given time.

7.1.6 Power Islands

A special case of power system topology is the *power island*, or an energized section of circuits separate from the larger system. An island would be sustained by one or more generators supplying a local load, at whatever scale. For example, in the event of a downed transmission line to a remote area in the mountains, a hydroelectric plant in this area might stay online and serve customers in its vicinity—an occurrence described by operators as “not by the book.”

Similarly, small-scale distributed generation such as rooftop photovoltaics combined with battery storage can in theory sustain local loads as a small island during a service interruption. This is increasingly common at the level of an individual house, where the solar inverter must first disconnect from the utility grid, but may then transition to powering the home from the battery in grid-forming mode (where the utility has neither control over nor liability for what happens). The

¹⁰ One side effect of protected network transformers is that they also prevent the export of locally generated solar power to the grid when that generation exceeds local load.

sequence of operations is of the essence in this scenario, to prevent the inverter from energizing distribution lines and neighboring customers.

Powering multiple customers on an islanded distribution circuit is more complicated. *Microgrids* (see Section 15.4) are grids of various sizes that can operate either separately in island mode, or connected to the rest of the grid. Crucially, microgrids rely upon a distribution infrastructure, especially protection systems and grounding, that has been specifically analyzed for safety and performance in both the islanded and grid-connected condition, with dedicated hardware for the transition between the two modes. Microgrids also require adequately sized resources (usually including a considerable amount of storage) and control systems that balance generation and load on the island and stabilize frequency and voltage. While the proliferation of clean, affordable distributed generation (Section 15.2) and storage (Section 15.3) along with growing interest in emergency preparedness and community resilience implies vast growth opportunities for microgrids, it is important to recognize the distinction between systems specifically designed or adapted for this purpose, and the majority of extant utility infrastructure.

Historically, *islanding* has not been routinely practiced or condoned by the U.S. utilities except temporarily, while restoring loads after a widespread outage. In these situations, power islands remain under the utility's control and are reconnected as quickly as possible, in a centrally orchestrated effort. Restoration scenarios typically involve larger grid segments powered by conventional power plants, although mobile equipment such as diesel generators may also be used. From the utility's standpoint, the key is that their crews have both visibility and control over any equipment that can energize power islands in these precarious transition states.

Reasons to avoid routine islanding of distribution circuits include both safety and liability. First and foremost, the safety of line crews could be jeopardized if they encountered a power island while expecting to find a de-energized circuit. A policy to permit *intentional islanding* beyond an individual customer's premises would mean that line crews must always treat distribution circuits as being energized—which is possible to do safely, but requires special precautions and equipment that costs time and money.

Furthermore, the ability of customer-owned generators or inverters on the island to maintain power quality (specifically, voltage and frequency) is not guaranteed. This could potentially cause problems for some customers with sensitive equipment, for which the utility may then be held liable without being able to do anything about it, or even knowing that a problem exists until it's too late.

These challenges notwithstanding, given the increasing interest in microgrids, along with growing concern about resilience in the face of extreme weather events, intentional islanding on utility distribution circuits is a subject of active research and development.

7.1.7 Loop Flow

In addition to protection, an important operational complexity introduced by a network structure is *loop flow*. Loop flow can arise whenever there is more than one path for the current to travel between two points in the system. The basic problem is that current flow cannot be directed along any particular branch in the network, but is determined by Kirchhoff's laws (see Section 2.3) and the relative impedances of the various branches. The concept is best explained with a toy example, shown in Figure 7.11.

How the power flows through a network tends to be of little interest until there is *congestion* or overloading of transmission lines, at which point it suddenly becomes critically important. In order for local transmission overloads to be relieved, operators need to know which generators could have their output adjusted so as to most effectively achieve a reduction in line flows. In a

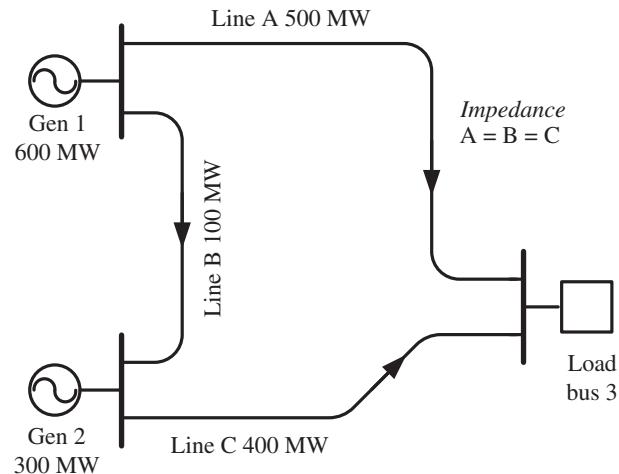


Figure 7.11 Loop flow.

competitive market, it may be desirable to allocate the available transmission capacity in some economic fashion. If an individual generator is to be allocated the rights to a certain power transfer along a particular link, it is vital to know how this generator's output contributes to the total flow on any given transmission path. While this contribution cannot readily be measured, it can be calculated or at least estimated by power flow analysis (see Chapter 12).

In the simplified situation shown in Figure 7.11, there is only one load, located at Bus 3, and two generators, Gen 1 and Gen 2. Suppose the load is 900 MW, of which Gen 1 supplies 600 MW and Gen 2 300 MW. What are the flows on the transmission lines A, B, and C? Although line A is the most direct path from 1 to 3, not all of the 600 MW will flow along A. Some will flow along the combination B-C, which constitutes another path from 1 to 3. The magnitude of this portion depends on the relative impedance of the path B-C as compared to A (see Section 2.2.2 about parallel circuits).

To make this example most transparent, let us suppose that the impedances of all three links A, B, and C are exactly the same (a highly idealized situation). The impedance of path B-C (in series) is therefore twice that of path A, and we would expect the current flowing through B-C to be roughly half of that flowing through A (see Section 1.2). The exact power flow solution will depend on the actual voltage at each bus. But a reasonable estimate for the sake of understanding loop flow is that roughly twice as much power flows through A as through B-C. If these are the only two paths from this generator to the load, and their total is 600 MW, then 400 MW flow through A and 200 MW through B-C.

Now there is an additional 300 MW supplied to the load at 3 from Gen 2. Again, while line C is the most direct path, some of the current will flow around the loop B-A. Given our previous assumption about the impedances, twice as much current (or power) flows through C as through B-A; thus, we have 200 MW through C and 100 MW through B-A, for a total of 300 MW.

The total flows on each line can now be calculated quite simply by taking advantage of the *superposition principle* (see Section 2.4), which allows us to consider each power source individually at first and then add the currents due to each source in each link. We must only be careful about the direction and whether the currents in fact add or subtract. On line A, power from Gen 1 and Gen 2 flow in the same direction, from 1 to 3. We can therefore add the currents (or power), and the total flow on A is 500 MW. Similarly, on line C, we have 200 MW from Gen 1 and 200 MW from Gen 2, each flowing in the direction from 2 to 3, for a total of 400 MW. But on line B, 200 MW from Gen 1 flow to Gen 2, whereas 100 MW from Gen 2 flow from 2 to 1. These line flows subtract, and we

have a net flow of 100 MW on line B from 1 to 2. We can do a reality check by confirming that the total power arriving at the load is indeed $500 + 400 = 900$ MW.

This discussion is intended to show that line flows in the presence of loops are not at all obvious, even for the simplest of cases. If there were more possible paths connecting the buses, the result would be accordingly more complicated and could hardly be calculated by hand. The example also illustrates that the effect of one generator's output may be to reduce rather than increase the flow on a given link. For example, suppose line B is overloaded and can only handle 90 MW, whereas there is plenty of capacity on lines A and C. If load at Bus 3 as well as generation at Gen 2 are now increased by 30 MW, the result is that the flow on B is reduced by 10 MW, saving the day. (The reader can verify that the new flows on A and C are 510 and 420 MW, respectively.)

Under some conditions, there may be a circulating current around a loop within the network that does not actually serve a load but contributes to energy losses. These cases show that designing and operating a power network for safety as well as efficiency requires modeling the power flow (Chapter 12) under a diverse set of operating conditions so as to avoid unpleasant surprises.

7.1.8 Reconfiguring the System

Operating a transmission and distribution system involves switching connections between and among high-voltage power circuits. Switching operations can be carried out remotely from the computer screen at a switching center through a *Supervisory Control and Data-Acquisition* (SCADA) system. In many distribution systems, however, switching is still performed manually by field crews per telephone instructions from operators at a distribution switching center.

Reasons for switching include contingencies, work clearances, service restoration following an *outage*, managing overloads, and enhancing system efficiency. For example, a contingency might be a fault on one line that has to be isolated from the system, and other connections need to be rearranged so as to redistribute the load of the lost line among them. Similarly, in the case of maintenance work or replacement of a line or system component, this component is electrically isolated (*cleared*) and the system “patched” around it. Clearances may also be issued for work other than on the power system components themselves, including any construction work that is sufficiently close to power lines that workers might accidentally contact them.

In the event of an outage, or service interruption to some number of customers, care is taken to follow a sequential procedure of restoring load. The idea is to reconnect sections of load one at a time in a given order so that each new load added does not jeopardize the stability of the remainder of the connected system. This restoration process involves opening and closing switches in order to divide loads into appropriately small sections, to connect them, and sometimes to transfer sections temporarily. Restoration after a widespread outage is usually the only time that utilities operate parts of the system as power islands. In this situation, local areas with generation and load are brought back on line and then synchronized with each other and reconnected.

At the transmission level, where outages are less frequent, switching operations isolate certain lines and equipment for maintenance without affecting loads. In some radial distribution systems, well-defined local blocks of load can be shifted from one circuit to another. For example, if one particular substation transformer threatens to overheat on a day of heavy demand, part of the load of the corresponding distribution feeder can be switched over to the adjacent transformer. This is easily done in a loop system by shifting the position of the “gap” or open switch between the two ends of the loop (see Figure 7.9).

Under extreme overload conditions, where there is either no available distribution capacity to reroute power, or if there is a shortfall in power generation, load may be *shed* or selectively disconnected. For example, if the systemwide generation is insufficient to meet demand, and consequently the system frequency drops below a specified limit, a number of customers will be disconnected for the sake of keeping the remainder of the system operational, as opposed to risking a more extensive failure of potentially much longer duration. To assure fairness in this procedure, customers are assigned *rotating outage block* numbers that appear on the electric bill for taking turns in being shed.

Finally, it is possible to redistribute loads in distribution systems for the purpose of *load balancing*, or equalizing the loads on distribution feeders or transformers in order to increase operating efficiency by minimizing losses. Because resistive power losses vary with the square of electric current ($P = I^2R$), they are minimized when the currents are evenly balanced among alternative lines and transformers.¹¹ To achieve this, some load may be switched over from one to another, less heavily loaded circuit. This procedure is not part of standard operations in the industry today, but has been proposed as an economically viable procedure in the context of automating distribution switching operations.

7.2 Qualitative Characteristics of Power Lines

7.2.1 Conductors

Conductors of overhead transmission and distribution lines typically consist of aluminum, which is lightweight and relatively inexpensive, and are often reinforced with steel for strength. *Stranded* cable is often used, which, as the name suggests, is twisted from many individual strands. At the same diameter or *gauge*, stranded cable is much easier to bend and manipulate. For underground lines, *cables* with insulation are used. Here heat dissipation is more of an issue, whereas weight is not. Copper is the material of choice for underground cables because, while it is more expensive, it has a lower resistance than aluminum. Low resistance is generally desirable for power lines to minimize energy losses, but also because heating limits the conductor's ability to carry current.

Recall from Section 1.2 that resistance is given by $R = \rho l/A$, where A is the cross-sectional area, l is the length of the conductor, and ρ (rho) is the resistivity (inverse of conductivity). The electrical resistance of a power line thus increases linearly with distance and decreases with the conductor cross-section (which, in turn, is proportional to the square of the radius or wire diameter). For the purpose of minimizing resistance, then, conductors should be chosen large. However, resistance must be weighed against other factors, including the cost of the conductor cable itself and its weight that needs to be supported by the towers or poles. Because even aluminum conducts so well, this trade-off comes out in favor of surprisingly slender lines, considering the amount of current and power transferred.

In practice, the current-carrying capacity of a transmission line may be limited by thermal expansion, and associated sagging. Innovative conductor materials reduce the coefficient of expansion, allowing a line to be operated at a higher temperature (and thus higher current) without violating minimum ground clearance. These are known as *high temperature low sag* (HTLS) conductors.

¹¹ Readers to whom this result is not obvious may wish to work through a numerical example: Consider two currents that add up to 10. In an unbalanced combination—say, 9 and 1—the squares of the currents add up to 82. In a balanced combination—that is, 5 and 5—the squares add up to only 50.

While resistance of lines is critical in the context of line losses, it is less important in the context of power flow and stability. This is because the overall impedance of most lines is actually dominated by their *inductive reactance*, to such an extent that it is sometimes appropriate to make an approximation where a line has zero resistance and only reactance (the *lossless line*). It is perhaps surprising that transmission and distribution lines should have any inductance at all, even though they do not resemble wire coils. Recall that inductance is based on magnetic flux lines linking a loop of wire (see Section 2.6). This notion extends to a straight wire, which can be considered an infinitely large loop, and the magnetic flux around the wire does link it. Since there is only a fraction of a turn in a straight line, this magnetic effect is quite weak. But it is cumulative on a per-unit-length basis, and with a conductor that extends over tens or hundreds of miles, it does eventually add up. There are two contributions to line inductance: the self-inductance, which is just a property of the individual conductor, and the mutual inductance, which occurs between the conductors of the three different phases.

Transmission lines have *capacitance*, too. It is a bit easier to see how two lines traveling next to each other would vaguely resemble opposing plates with a gap in between. In fact, there is also capacitance between a conductor and the ground. Because the lines are small and the gap wide, the capacitance tends to be fairly small. Capacitance is especially important, though, for *coaxial* cables where one conductor surrounds another with insulation in between. Coaxial cables are used where simplicity and compactness matters; for example, on underground or undersea cables for d.c. transmission.

In describing transmission-line parameters, the inductance is considered to be in series along with the resistance, and the capacitance in parallel (a *shunt element*); we save the details for Chapter 9. Qualitatively, we note here that a line is characterized by an equivalent resistance, inductance, and capacitance on a per-meter or per-mile basis, since all the above quantities are additive over the length of the line. In addition, a shunt conductance would account for electrical discharge through the air (known as *corona losses*), but this quantity should be very small and is often neglected.

Figure 7.12 and Table 7.1 give some examples for physical dimensions and electrical properties of transmission lines. Note that the inductance, since it is modeled in series, corresponds to a relatively small impedance. The capacitance, by contrast, is modeled in parallel and therefore corresponds to a large impedance (since the larger a parallel impedance, the smaller its impact).

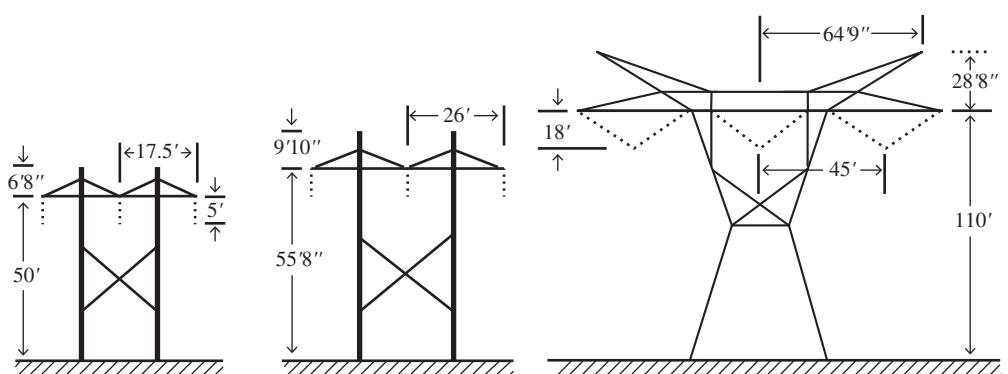


Figure 7.12 Sample transmission-line dimensions. Source: From EPRI, 1977.

Table 7.1 Sample transmission line data.

	Line Voltage (kV)		
	138	345	765
Conductors per phase	1	2	4
Number of strands aluminum/steel	54/7	45/7	54/19
Diameter (in.)	0.977	1.165	1.424
Conductor geometric mean radius (ft)	0.0329	0.0386	0.0479
Current-carrying capacity per conductor (A)	770	1010	1250
Geometric mean diameter phase spacing (ft)	22.05	32.76	56.7
Inductance ($\text{H/m} \times 10^{-7}$)	13.02	9.83	8.81
Inductive reactance X_L (Ohms/mi)	0.789	0.596	0.535
Capacitance ($\text{F/m} \times 10^{-12}$)	8.84	11.59	12.78
Capacitive reactance X_C (MOhms/mi)	0.186	0.142	0.129
Resistance (Ohms/mi)	0.1688	0.0564	0.0201
Surge impedance loading (MVA)	50	415	2268

Source: Source: Bergen et al. 2000/with permission of Pearson Education.

7.2.2 Bundled Conductors

Conductors on transmission lines—especially high-voltage, high-capacity lines—are sometimes *bundled*, meaning that what is electrically a single conductor is actually composed of two, three, or four wires a few inches apart, held together every so often with connectors known as *conducting frames*. There are several reasons for bundling conductors: increasing heat dissipation, reducing corona losses, and reducing inductance.

The first of these reasons is straightforward: by dividing a conductor in two (or more), the cross-sectional area (and thus the resistance and the weight) can be kept the same while increasing the surface area. This larger surface area allows the conductor to more effectively radiate heat off into the surrounding environment. Thus, for the same amount of power dissipated (at the same current), the conductor's equilibrium temperature will be lower.

The second major reason to bundle conductors is to reduce line inductance. This phenomenon is much less intuitive, and discussed in Section 9.1. In sum, careful analysis of the magnetic flux both outside and within a conductor shows that the inductance is less for a bigger wire diameter, and less if the three phases are closer together. Changing a single wire into a bundled conductor makes it resemble a wire of larger diameter as far as the magnetic field is concerned; reducing the magnetic field or flux linkage in turn reduces the line's inductance. The effective conductor size approximated by a set of bundled conductors is described in terms of a *geometric mean radius* (GMR), a quantity listed with sample values in Table 7.1.

Another advantage of increasing surface area by bundling conductors is reduced *corona losses*. The corona results from the electric field that surrounds the conductor at high voltage. Similar to the top of a *Tesla coil*,¹² though fortunately much less intense, microscopic arcs occur between the conductor surface at high potential and ionized air molecules in the vicinity. The high frequency in a Tesla coil results in an arc “boring” its way into the air much farther than one would expect

12 A device invented by Nikola Tesla (1856–1943) that produces very high voltages and lightning discharges by transforming a rapidly alternating current.

based on the ionization potential of air (see Section 1.1.4). This is because, unlike the case of a static electric field, a fresh batch of air molecules is ionized each time the polarity reverses, so that the power carried away from the conductor increases with frequency.¹³ Since there are other ways for ionized air molecules to travel, the corona effect is sensitive to weather and does not entirely vanish for direct current.

At 50 or 60 Hz, though much less potent than a Tesla coil, a corona of tiny arcs that discharge into the air produce an audible crackling sound around high-voltage a.c. equipment, especially for voltages in the hundreds of kilovolts. Because the arcs are so small, they are not visible to humans even at night. Yet there is a measurable energy loss associated with what is in fact a small electric current flowing to ground through the air. The power associated with this current is the corona loss. When the surface area of the conductor is increased, the electric potential or surface charge density is spread out more, reducing the electric field strength. This in turn reduces the formation of arcs, thereby reducing corona losses.

7.2.3 Towers, Insulators, and Other Components

The poles or towers that support overhead transmission and distribution lines are usually made of wood or, for the larger transmission towers, metal. Designs used by different utilities vary depending on line voltage, conductor size and weight, terrain, aesthetic preferences, and tradition. Ideally, the distance between conductors is maximized while using a minimum amount of materials for tower construction. In any case, towers have to be made tall enough so that there is sufficient clearance between the hanging conductor and the ground or any other objects that may threaten to come into contact with an energized line. Of course, clearance also has to be maintained between the conductors and the towers. Depending on the line voltage, there are standard design criteria for minimum safe clearance, which have to take into account the *sagging* of lines due to thermal expansion under high current.

In order to achieve a given clearance, there is an engineering trade-off between making the towers tall versus using more towers at closer spacing. Sometimes the separation between towers is constrained; for example, when crossing a body of water. Here, the towers have to be made extra tall in order to allow for sufficient clearance.

Insulators serve to electrically separate the conductor from the tower. They are made in carefully designed, rounded shapes consisting of one or several *bells* made of a nonconducting ceramic or plastic. The surface of the bells is round and smooth to minimize the potential for arcs to form. Dividing a long insulator into a string of individual bells is more effective than a single cylindrical shape, because it provides more surface area. This area helps spread out surface charge and discourages a current from creeping along the surface, especially when the insulator is wet. The connection point between high-voltage conductors and insulators sometimes also features *corona rings*, which increase the surface area of the metal conductor and thus reduce the electric field and arcing potential.

The length of insulators, or number of bells, is roughly proportional to the line voltage, which corresponds to the potential difference that the insulator has to maintain.¹⁴ As a rule of thumb,

¹³ The empirical relationship between power and frequency, roughly linear over some range, is known as *Peek's formula*.

¹⁴ To be precise, the insulator must be able to sustain the maximum instantaneous line-to-ground voltage (given that the tower or pole is at ground potential), whereas the "line voltage" is conventionally quoted as the line-to-line rms value (see Section 3.1.3). The maximum is greater than the root mean squared (rms) value by a factor of $\sqrt{2}$, while the line-to-ground voltage is less than line-to-line by a factor of $\sqrt{3}$. The insulators on a 230-kV line, for example, must sustain an instantaneous potential difference between conductor and tower of $230\text{ kV} \cdot \sqrt{2}/\sqrt{3} = 188\text{ kV}$ under normal operation.

each one of the typical bells contributes to insulate against a voltage on the order of 10 kV. Thus, counting the number of bells on the insulators can give a rough idea of the voltage on a transmission or distribution line. Single bells are used for primary distribution below or around 10 kV.

Most often, transmission lines hang down from the tower on a single insulator, but sometimes one sees configurations with horizontally extended insulators and the conductor describing a semicircle underneath. These are the places where horizontal tension is applied to the conductors. In order to create a reasonable clearance and not let the wires droop down, the amount of tension is several times greater than the actual weight of the conductor.¹⁵

Owing to their shape and conductivity, metal transmission towers are a likely target for lightning strike. The electric current associated with lightning can usually travel through a metal tower into ground without doing any damage. However, one wants to prevent lightning from traveling along the actual conductors, where it would cause severe voltage fluctuations and potential equipment damage. The metal top of a transmission tower serves to attract lightning and direct it straight down into the tower to ground. In areas where lightning is common, wooden distribution poles are specifically equipped with metal lightning arresters. In addition, it is common to connect metal towers with a ground wire—usually a small diameter wire at the top of the towers—that is electrically separate from the power circuit. In case one tower experiences a change in potential due to lightning, the grounding wire provides a path for current to flow and the potential to equalize among neighboring towers. This minimizes the danger of an arc jumping across an insulator between a tower and conductor and lightning current flowing along the conductor.

An unrelated type of item one sees on transmission lines are *vibration dampers* attached to conductors, usually near the tower. Their function is to reduce the amount of swinging and vibration of the conductor in the wind, by changing its mechanical resonant frequency. There are also the large red and white plastic balls which, as most people guess correctly, are intended for airplane and helicopter pilots to see.

Occasionally, transmission towers rotate or *transpose* the positions of the three phases, as in Figure 7.13. The purpose of transposition is to keep the mutual inductance roughly equal on all three phases, which is desirable for purposes of balancing the load (not to mention keeping up with the engineering analysis). Arranging the three conductors symmetrically as an equilateral triangle solves the problem, but is not always practical, depending on the tower or pole design. If three conductors are in a row, there is an asymmetry because the one in the middle is closest to both other conductors. Over a long distance, therefore, the conductors are transposed every so often, allowing each individual phase to travel roughly the same distance in each of the three positions.

7.2.4 DC Transmission

The bulk of this chapter and Chapter 8 is devoted to a.c. transmission, for two reasons. The vast majority of extant power transmission and distribution infrastructure operates on alternating current. Also, since the analysis of a.c. systems is more complex—literally and figuratively—and involves more variables, a.c. transmission naturally takes many more pages to explain.

¹⁵ Readers with some background in physics may enjoy the classic problem of deriving the parametric equations that specify the shape described by the conductor suspended under tension (no, it is not exactly a parabola). (Hint: Consider the angle relative to horizontal that the conductor makes at any given point. Note that any given small segment of conductor is not accelerating (the forces on it are equal in all directions), and that the force pulling it toward the bottom of the curve is determined by the mass below it.)

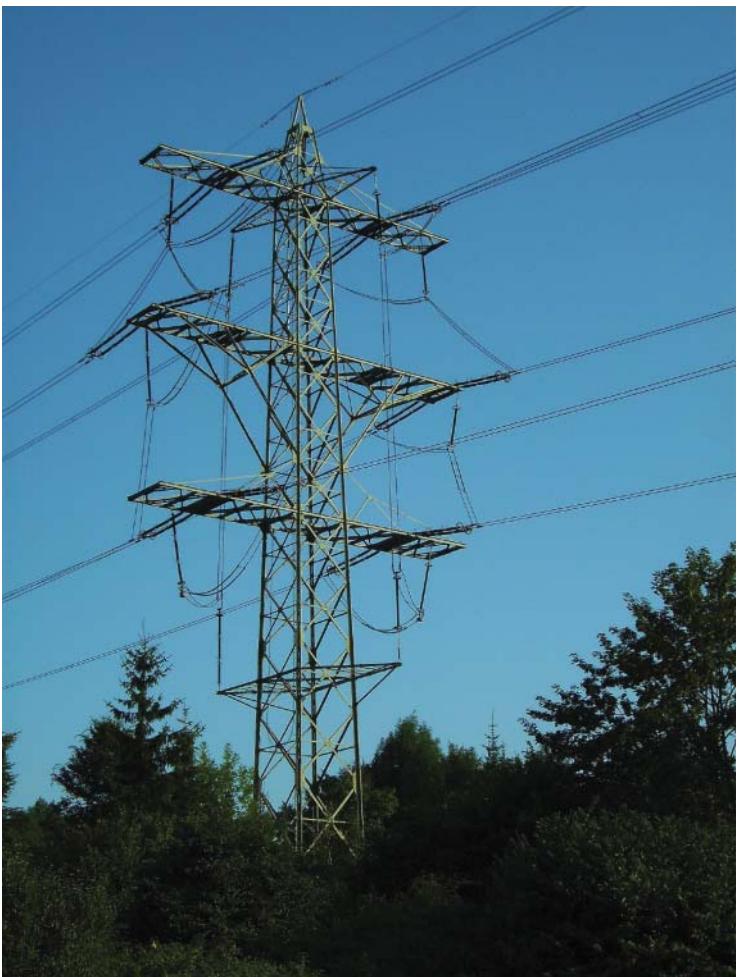


Figure 7.13 Transposition tower for a 380-kV line in Germany, carrying two three-phase circuits. For example, the middle phase at the front left is moved to the top position on the right. Also discernible on the photo are bundled conductors with two subconductors per bundle, and corona rings at the ends of the insulators. Source: Zonk43/Wikimedia Commons/Public domain.

But as of the early 2020s, there is growing interest and investment in high-voltage direct-current (HVDC) transmission, suggesting that the topic will be of substantial interest to future power engineers. This section offers a brief introduction to the main characteristics of d.c. lines.¹⁶

First of all, the HVDC label emphasizes the crucial contrast between contemporary d.c. transmission and historical d.c. infrastructure that was limited by low voltage. Recall from Section 7.1.1 that the “Battle of the Currents” was won by alternating current because transformers, which only work with a.c., were the only practical technology at the time to step voltages up and down. D.C. transmission in Edison’s time was therefore synonymous with inefficient, low-voltage transmission.

Today’s HVDC lines benefit from solid-state technology that readily converts power between any form of a.c. or d.c. at different voltage levels (see Chapter 14). D.C. transmission voltages are

¹⁶ The abbreviation format a.c. and d.c. or AC and DC is a matter of stylistic preference. This book mostly uses the lowercase format, although in the context of high-voltage transmission, AC and DC are more common.



Figure 7.14 The ± 500 kV Pacific DC Intertie (PDCI) and a three-phase 230 kV a.c. transmission line near Bishop, California.

often greater than a.c. lines, in the hundreds of kilovolts, but ultimately limited by practical design constraints and cost trade-offs for all the components (lines, towers, and the high-voltage side of converter stations) that must be properly insulated and protected against arcing faults.

A d.c. transmission line is easy to recognize in that it has two, not three conductors per circuit. Figure 7.14 shows side-by-side d.c. and a.c. lines. The ± 500 -kV Pacific DC Intertie (PDCI) shown here connects two converter terminals, Celilo in the Pacific Northwest and Sylmar in Los Angeles, and serves to stabilize north–south power flow through Oregon and California while adding 3100 MW of rated power transfer capacity to the backbone of the West Coast transmission grid. The PDCI has no other electrical connections along the way, but is seen here enjoying the company of a 230-kV a.c. line for a stretch of its scenic path through the Owens Valley.

Since they are augmenting an existing a.c. network, d.c. lines typically just connect two points, rather than branching out or serving local loads. Converter stations at each end interface with the rest of the grid through bi-directional conversion between a.c. and d.c. The legacy technology is based on *thyristors*, while modern stations use *modular multi-level converters* (MMCs) that allow for more control on the a.c. side (see Chapter 14). Because of the fixed cost associated with building converter stations, and because the cost per mile for a given power transfer capability is less for d.c. than a.c. transmission, the economics generally favor longer HVDC lines (on the order of hundreds of miles).

Long transmission distances also capitalize on a specific advantage of HVDC lines: namely, that they are not subject to a stability limit like long a.c. lines (see Section 7.3.2). This is because the d.c. line imposes no physical constraints on the relationship between the voltage phase angle at either end. For an a.c. line, that phase angle separation is physically tied to the voltage, the inductive reactance, and the amount of power transferred (see Eq. (7.1)). Over long distances, an a.c. line's

inductive reactance X necessarily increases, meaning that less power can be transferred without forcing a voltage phase angle separation that could cause dangerous instabilities or oscillations. By contrast, each converter at the end of a d.c. line independently ties into the a.c. side with the appropriate voltage magnitude and angle for each location. In fact, modern control technology at the converter station can help drive and stabilize this voltage.

HVDC lines have other technical advantages. Since the voltage is essentially constant, there is no factor of $\sqrt{2}$ between the peak and average (rms) value, which means better utilization of the asset that must be built to withstand the peak voltage.¹⁷ Better asset utilization also applies to current in that there is no reactive power transfer on a d.c. line, and the useful power transferred is given by the entire product of current and voltage. In addition, HVDC lines tend to have less *corona losses* (introduced in Section 7.2.2) than a.c. lines, as the corona effect decreases at lower frequency. However, it also increases dramatically with voltage, so that ultra-high voltage d.c. lines still contend with the phenomenon.

Besides economic long-distance transmission and improving wide-area stability, HVDC lines are sometimes used to transfer power between asynchronous a.c. networks, where asynchronous means that there is no particular relationship between the frequency or timing of the waveform. This allows sharing resources without coupling the systems operationally through the same frequency control. Possible reasons for adjacent systems that otherwise cooperate to remain asynchronous with each other include large geographic distances (as between the Eastern and Western U.S.), different nominal frequencies (as in Japan, which for historical reasons uses 50 and 60 Hz in different parts of the country), or different frequency regulation standards (as between East and West Germany after the 1989 reunification).

Finally, HVDC cables are used for undersea power transmission. In this application, direct current entirely avoids the problem of capacitance posed by insulated high-voltage cables, which becomes significant over medium distances (on the order of tens of miles). Undersea transmission has been strategically important especially for interconnection of the European grid, as well as the development of offshore wind generation.

7.2.5 Superconducting Transmission

Superconductivity obviates many constraints and concerns about transmission lines, most notably thermal loading limits and line losses. These constraints matter not just in terms of making transmission marginally more economical, but they could be strategically important, in that newly sited transmission projects are often a prerequisite for developing renewable generation resources that are distant from load centers. Increasing the load carrying capability of a transmission path by an order of magnitude through superconducting technology would imply a greatly reduced need for new conductors or new rights of way.

As of this writing, “high-temperature” superconductivity is fairly well-understood at temperatures above the boiling point of liquid nitrogen (-196°C , see Section 1.1.4), while ambient or room-temperature superconductivity remains elusive for practical purposes. This means that sophisticated refrigeration is required to maintain the superconducting state over a large mass of material and over time. To date, high-temperature superconducting transmission has been deployed in niche applications over short distances. Clearly, the capacity and efficiency gains have to justify the expense (in both money and energy) of cooling equipment alongside a

¹⁷ See Section 3.1.3 for the root-mean-square value. The same math applies to current, but the design constraint for maximum current rating is already based on the rms value because it relates to cumulative heating, so there is no specific advantage in d.c.

superconducting transmission line. Approaches to reduce this expense include replacing some of the refrigeration or subcooling with evaporative cryogenic cooling.¹⁸ Note also that active refrigeration implies some risk exposure to equipment failure.

Although a superconducting transmission line has no resistance, it still has reactance, since the relevant electric and magnetic fields are external to the conductor. Therefore, practical superconducting long-distance transmission lines will almost certainly use direct current, as their loading capacity would otherwise be limited by inductive reactance. It also stands to reason that cooled, insulated cables would be more practical to bury underground rather than stringing overhead. Finally, note that high current allows the use of lower voltage levels, which would tend to reduce the costs of lines as well as HVDC converter stations.

7.3 Loading

7.3.1 Thermal Limits

Distribution lines and short to medium transmission lines are limited in their capacity to transmit power by resistive heating. It is thus the magnitude of the current, continuing over time and increasingly heating the conductor, that limits the loading; this is the *thermal limit*. As the conductor heats up, it stretches from thermal expansion, and the line sags. If it sags too far, the distortion becomes irreversible. This is bad for the conductor and may violate the clearance requirement. Also, the resistance of the conductor will increase with temperature. This is generally a small effect, but it does eventually become noticeable, especially since increasing resistance in turn will increase heating, not to mention losses. In the worst-case scenario—say, if a fault did not get cleared—a conductor can actually melt off the pole.

Because conductor temperature is the real limiting factor, the *rating* of a line that states the amount of current it can safely carry is an approximation based on assumptions about the weather. If it is cold and windy, the line can dissipate more power while remaining at the same temperature. General practice among transmission and distribution engineers has been to err on the safe side and rate lines conservatively for hot weather with no wind. However, in the interest of improved asset utilization, it has become common to adopt variable ratings that take into account ambient conditions. A relatively crude approach is to have a summer and a winter rating. Of course, this does not help very much if load peaks and transmission congestion occur in the summer. *Dynamic ratings* are ones that are updated more frequently with current weather information.

Besides the ambient conditions, it is also important to consider the loading history of the line, since heating occurs gradually. In this vein, lines often have a *normal* and an *emergency rating*. The emergency rating may be good for a number of minutes or hours, which may be just long enough to sustain the load in case of a contingency (such as the loss of another line).

Since current is the key variable for heating, line ratings are generally expressed in terms of *ampacity*, or current-carrying capacity in amperes. This ampacity is independent of voltage. Thus, the amount of power that can actually be transmitted by a given conductor depends on the operating voltage. Note that when the current is translated into power, this means apparent power in volt-amperes (VA), not real power in watts, since only the total magnitude and not the phase of the current matters for heating purposes.

¹⁸ Fortunately, nitrogen is about as innocuous as it gets for venting a gas into the atmosphere.

Heating also limits the operation of transformers at high current. The situation is slightly more complicated because the amount of energy dissipated within the transformer core depends to some extent on voltage (and on a.c. frequency, though this is unimportant here), but the key concern is again temperature. Although transformers are not as exposed to the elements as power lines, environmental factors such as ambient temperature, wind, and rain do play a role, as does the circulation of oil or coolant. Transformer ratings are given in terms of apparent power (kVA or MVA).

Because of the importance of environmental factors as well as the time dimension in overheating equipment, any thermal rating is inherently approximate. As mentioned in the context of generators, there has been some historical and cultural trend from conservative toward more exact ratings, a trend that surely is not unique to the electric power industry. Depending on the vintage and design of an individual piece of utility equipment, its official nameplate rating may or may not coincide with an engineer's or operator's judgment of what loading it could actually withstand at a given time without damage. Variable ratings can be understood as an effort to formalize this judgment while deriving maximum economic benefit from the extant hardware.

7.3.2 Stability Limit

Aside from heat, another type of limit on power transmission is sometimes important for longer transmission lines: the *stability limit*. Most often, “stability” in this context is taken to mean *angle stability*, although *voltage stability* is sometimes relevant as well. Both types of stability are discussed in more detail in Section 13.4; here, we briefly introduce the concept and the practical relevance of a stability limit.

Angle stability relates to maintaining the feedback between generators on either end of the line that keeps them locked in synchronicity, by making each generator push harder if it tries to speed up and less hard if it tries to slow down. Casually speaking, transmitting power along a line requires that a generator on the sending end pushes harder than one on the receiving end. Accordingly, the sending generator has a *power angle* δ that is somewhat ahead of the receiving generator. This power angle indicates the exact timing of the generator electromotive force (*emf*) or voltage pulse in relation to the voltage maximum of the system.

As derived in Section 13.5, the amount of real power transmitted on an ideal, *lossless* line (with no resistance, only reactance) is given by the equation

$$P_{12} = \frac{|V_1||V_2|}{X} \sin \delta_{12} \quad (7.1)$$

where δ_{12} is the difference in power angles between the sending and receiving end of the line,¹⁹ V_1 and V_2 are the voltage magnitudes at either end, and X is the reactance of the line in between. Figure 7.15 illustrates the relationship from Eq. (7.1).

In order to transmit a large amount of power on a given line, it is necessary to increase the angle difference δ_{12} . However, this difference cannot be made arbitrarily great. Not only does Eq. (7.1) impose a theoretical limit on real power transfer (at $\delta_{12} = 90^\circ$), but well before that hard limit is reached, there is a practical limit associated with an increasing risk of instability. In essence, this is because the stabilizing feedback between generators at either end of a transmission line becomes less effective for large δ_{12} , and they are at greater risk of losing synchronism (and disconnecting to potentially cause a blackout) due to some small disturbance.

For short lines, the reactance X in the denominator is usually small, so that a small δ_{12} still results in a large amount of power transmitted. Consequently, if one were to apply the maximum

¹⁹ In Chapter 12, the same angle is called θ .

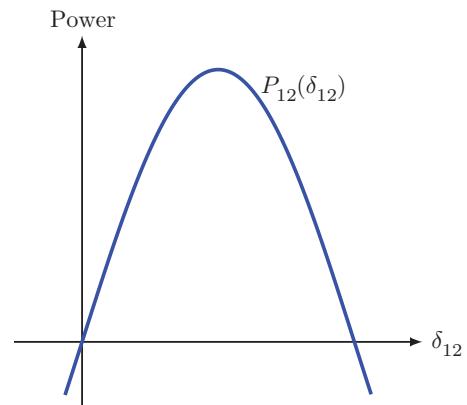


Figure 7.15 Real power flow as a function of voltage phase angle difference.

permissible δ_{12} to such a line, the power transmitted could easily exceed the line's thermal capacity. For long lines, however, the reactance becomes more significant. In this case, a dangerous δ_{12} well may be reached before the thermal limit of the line. Hence, a more stringent limit on power transmission called the *stability limit* is imposed. While the thermal limit is expressed in terms of either current (amps) or apparent power (MVA), the stability limit has units of real power (MW).

7.3.3 Surge Impedance Loading

Figure 7.16 shows the stability limit and thermal limit as a function of length for a hypothetical transmission line. The main idea in the diagram is that for a short line the thermal limit applies, whereas beyond a certain length the stability limit becomes more constraining. The value of δ_{12} for the angle stability limit is based on engineering judgment. The label P_{12}/P_{SIL} is a measure of the real power transmitted between the two ends of the line, expressed as a fraction of the *surge impedance loading* (SIL), which is a fixed characteristic of a given line.

Specifically, the SIL represents the amount of power transferred when the load connected at the end of the line matches the *characteristic impedance* of the line itself. This characteristic impedance is given by the relationship between inductive properties (captured in the series impedance) and capacitive properties (captured in the shunt admittance) on the line. When resistance is negligible compared to reactance, as we often assume it is for transmission lines, the characteristic impedance is called the *surge impedance*.

The significance of the surge impedance in telecommunications is that a signal can be transmitted with minimal loss if whatever is connected to the end of a line matches the line's surge

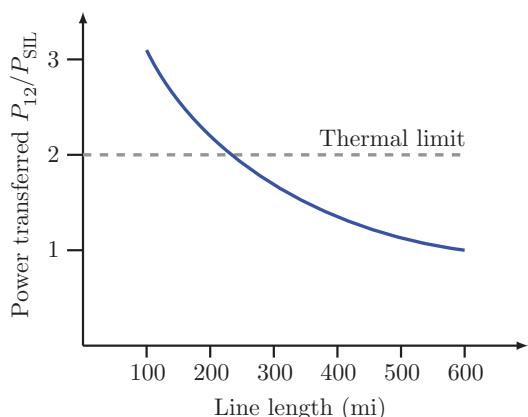


Figure 7.16 Thermal and stability limits for a hypothetical line.

impedance, because there is no reflection at the boundary (see Section 9.3.8 for more about wave propagation on transmission lines). In the grid context, it is more common to speak of a line's SIL in units of power. As shown in Eq. (7.4), SIL is given by the square of transmission voltage divided by the surge impedance. Table 7.1 list some sample SIL values in units of MVA.

The surge impedance load does not measure a line's power carrying capacity, but rather states the amount of real power transmission in the situation where the line's inductive and capacitive properties are completely balanced. To system operators, this provides a benchmark for reference: if the power transmitted along a line (at unity power factor) is less than the SIL, the line appears as a capacitance that injects reactive power (VARs) into the system; if transmitted power exceeds the SIL (the more common situation), the line appears as an inductance that consumes VARs, and thus contributes to reactive losses in the system.

The fact that the inductive property dominates at higher loading makes some intuitive sense because the line's reactive power consumption is a function of the line current, whereas the capacitive property of injecting reactive power is a function of the voltage at which the line is energized. Thus, the capacitive property will be largely indifferent to the load, while the inductive property will become increasingly pronounced at higher load.

In equation form,

$$Q_{\text{loss}} = I^2 X_L \quad \text{and} \quad Q_{\text{prod}} = \frac{V^2}{X_C} \quad (7.2)$$

From these equations we can also see that the transmission line "consumes" and "produces" equal amounts of reactive power (that is, $Q_{\text{loss}} = Q_{\text{prod}}$) when $X_L X_C = V^2/I^2$. Taking the square root of that equation puts it in units of impedance, or a ratio of voltage to current. This gives the characteristic impedance Z_C of a transmission line, or *surge impedance* specifically for a lossless line. Neglecting resistance and substituting the explicit inductance and capacitance from $X_L = \omega L$ and $X_C = 1/\omega C$, we obtain the surge impedance

$$Z_C = \frac{V}{I} = \sqrt{\frac{L}{C}} \quad (7.3)$$

Note that the a.c. frequency ω cancels in the expression, which tells us that the surge impedance is truly a characteristic of the line itself and not the a.c. environment in which it is being operated.

The SIL, or power transferred when the load impedance matches the line's surge impedance, is then given by

$$\text{SIL} = \frac{V^2}{\sqrt{L/C}} \quad (7.4)$$

These relationships are presented with more formal context in Section 9.3.10. Although derived for a lossless line, the SIL is still a useful approximation for realistic lines.

7.4 Voltage Control

Voltage in power systems is controlled both at generators and on location throughout the transmission and distribution system. As described in Section 10.4, the voltage at a generator terminal or *bus* is controlled by the excitation or rotor field current, which determines the rotor magnetic field strength and thus the magnitude of the induced *emf* in the armature. Generator voltage is directly linked to reactive power generation; the two variables cannot be controlled independently of each other. Since various generators may be producing different amounts of real and reactive power, their bus voltages vary slightly from the nominal voltage that is the same for all.

This variation is part of a subtle profile of voltage levels that rise and fall on the order of several percent throughout any interconnected power system. Such a profile exists separately from the

order-of-magnitude voltage changes introduced by transformers. For example, within a network that is nominally operating at 230 kV, the actual voltage at different locations may vary by thousands of volts (1% being 2.3 kV or 2300 V).²⁰ The exact voltage level at each location depends mainly on two factors: the amount of reactive power generated or consumed in the vicinity, and the amount of voltage drop associated with resistive losses.

In radial distribution systems, the voltage-drop effect dominates. Here the voltage simply decreases as one moves from the substation (the power source, in effect) out toward the end of a distribution feeder. This change in voltage is known as the *line drop*. The line drop is described by Ohm's law, $V = IZ$, where I is the current flowing through the line, Z is the line's impedance, and V is the voltage difference between the two ends. Ohm's law also shows us that the line drop depends on the connected load, since a greater power demand implies a greater current. While the line impedance stays the same, the voltage drop varies in proportion to the load.

In practice, the voltage drops in distribution systems are quite significant, especially for long feeders. Recognizing that it is physically impossible to maintain a perfectly flat profile, operational guidelines in the United States generally prescribe a tolerance of $\pm 5\%$ of the nominal voltage, while some countries use $\pm 10\%$. This range applies throughout transmission and distribution systems, down to the customer level. For example, a customer nominally receiving 120 V should expect to measure anywhere between 114 and 126 V at their service drop.

Figure 7.17 illustrates the problem of voltage drop along a radial feeder. If the feeder is very long, the voltage drop may exceed the window of tolerance, so that if the first customer is receiving no more than 126 V, the last would receive less than 114 V. In order to maintain a permissible voltage level along the entire length of a feeder, it may therefore be necessary to intervene and boost the

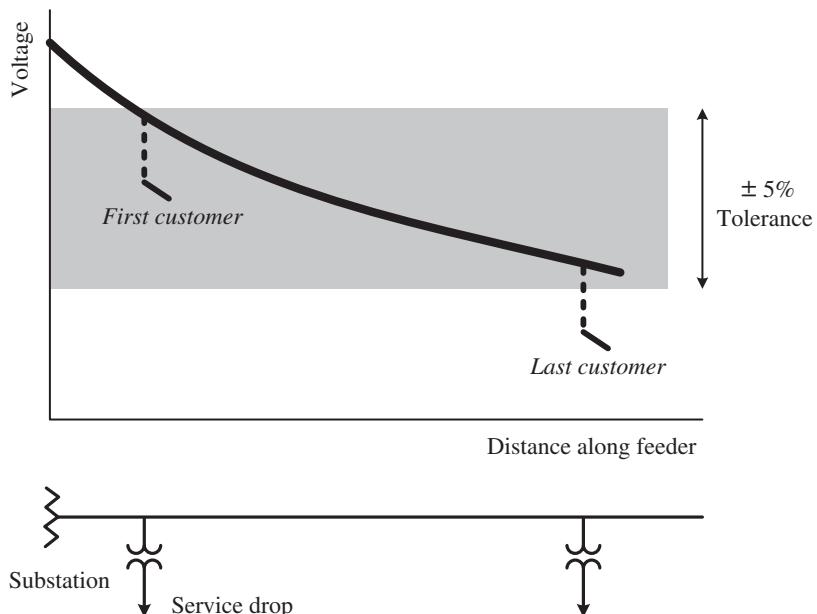


Figure 7.17 Voltage drop along a distribution feeder.

²⁰ Birds sitting on power lines are famously unharmed because their entire bodies are at the same voltage as the line, with no current flowing through them. In fact, this is only because their feet are so close together. If a giant bird could straddle across many miles to touch two ends of the same line, it would get terribly shocked.

voltage somewhere along the way. Furthermore, because the voltage drop varies with load, this boost may need to be adjusted at different times.

7.4.1 Tap Changers

There are two common methods for controlling or supporting voltage in transmission and distribution systems, based on very different physical mechanisms: transformer taps, and reactive power injection.

Adjustable transformer taps provide for a variable turns ratio, and thus a variable amount of voltage change effected by the transformer. The tap is simply where the conductor connects to the transformer coil on the secondary or load side. By moving the tap up or down, the effective number of turns of that transformer winding is changed. This mechanism is called a *load tap changer* (LTC), or on-load tap changer (OLTC) to indicate that the load is not interrupted during the tap transition. An LTC has some number of discrete settings that adjust according to the loading conditions downstream. A standard design has 16 steps up and down from neutral, to achieve a range of $\pm 10\%$ of voltage. Tap changing transformers of various sizes are used throughout transmission and distribution systems. A typical use would be at the substation where distribution circuits originate, also called the *feeder head*.

A related device that might be installed somewhere midway on a feeder rather than at the substation is called a *voltage regulator*, illustrated in Figure 7.18. This would be necessary if the voltage drop on a long feeder is so large that no setting at the feeder head could provide the appropriate voltage all the way from beginning to end. Voltage regulators look like tall transformers on distribution poles, often with large fins for heat dissipation, but with no secondary lines going out to customers; they are simply transformers between two segments of the same line, with the turns

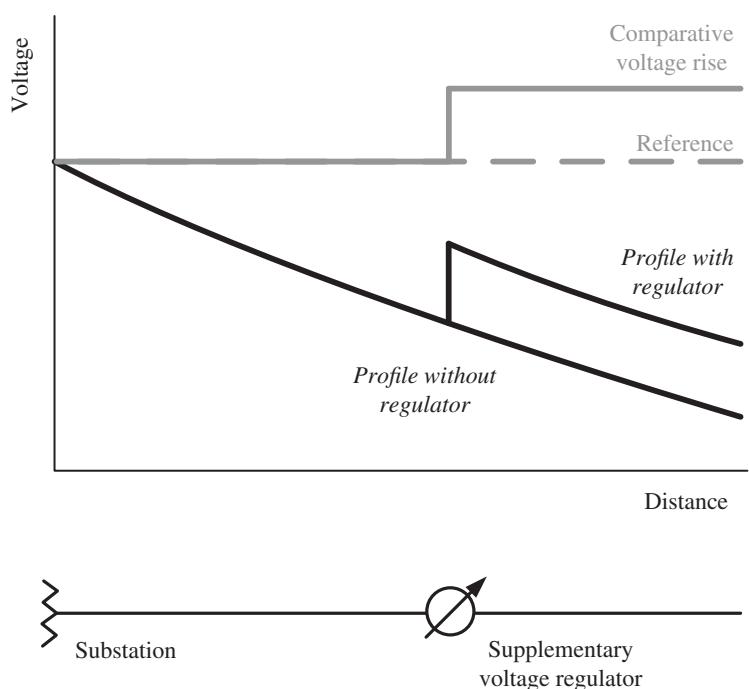


Figure 7.18 Effect of a line voltage regulator on the voltage profile of a radial distribution feeder.

ratio adjusted to boost the voltage just enough to compensate for line drop. A common design principle for a voltage regulator is an *autotransformer* (Section 8.4).

7.4.2 Reactive Compensation

Reactive power affects voltage through a very different mechanism than transformer turns ratios. For reasons discussed further below, injecting VARs anywhere in a transmission and distribution network tends to raise the voltage magnitude near the injection location. Reactive power can be sourced from *capacitors*, *synchronous condensers*, or *static VAR compensators* (SVCs).

Synchronous condensers are just rotating synchronous machines (i.e., motor/generators, see Section 10.3) operating at zero real power output. In principle, any conventional generator facility can be used or adapted for this purpose. If the plant is also producing real power, this simply limits its available VAR capacity to stay within the overall machine rating (see Section 10.5). A synchronous condenser built strictly for the purpose of VAR support does not require any of the parts of a power plant involved with supplying real power, as it will just consume a modicum of real power in operation to make up for losses. Synchronous condensers are the legacy technology for providing VARs at the transmission level, in areas where regular generation facilities are insufficient to manage voltage.

SVCs are a modern solid-state alternative based on semiconductor technology, with no moving parts. Specifically, SVCs use *thyristor* controlled components (see Section 14.3.3) that offer a near-instantaneous response to changing line voltage. An SVC installation may include a combination of thyristor-switched reactors, thyristor-switched capacitors, and mechanically switched capacitors (which are less expensive, but slower to respond). The function of the reactors is opposite that of the capacitors (i.e., they consume VARs so as to lower the voltage), which can be desirable to limit voltage rise on a transmission line at low load or when it is first energized.

All of these devices are similar in that they appear to the system as a capacitance when called for, with the result of boosting the local voltage magnitude. While the VAR output of SVCs and synchronous condensers can be continually adjusted, capacitor banks (where the term “bank” refers to all three phases) are controlled simply by being switched into or out of the system. Capacitors come in a wide range of sizes and are common at the distribution level.

Mechanically switched capacitor banks can be automatically controlled, either by sensing local variables such as voltage or current, or very simply by the time of day, which might be sufficiently well correlated with load. Any such capacitors are usually connected in parallel with the load; a parallel capacitance is also known as a *shunt capacitance*.²¹

Series capacitance is used in some specific applications, usually on transmission lines. The disadvantages of series compensation include the fact that the capacitor will be subject to the entire line current including possible fault currents, and that bypass switches are required in order to remove the capacitor from the circuit for maintenance. Series capacitors are also more likely to create unintentional resonance conditions with inductances. On the plus side, the effect of series capacitors is naturally correlated with load, since the amount of reactive power injected varies directly with current as I^2X —whereas the effect of shunt capacitors is given by V^2/X , which unfortunately

²¹ This makes sense if we consider that one would generally wish to make only a small adjustment, and ask what size capacitor would be needed. In the series case, we can think of the capacitor as replacing a short circuit, so in order to add only a small impedance, the capacitor needs to be huge. In the parallel case, the capacitor is effectively replacing an open circuit (two previously unconnected parts of the circuit), and adding a small capacitor means adding a small amount of admittance (or lowering the impedance by a small amount).

reduces their contribution under low voltage conditions when they are needed most. However, shunt capacitors are far simpler and more economical, and thus the norm for distribution systems.

The voltage rise due to reactive compensation can be most easily explained in terms of reducing the voltage drop on a radial line with an inductive load at the end. Compensating for that inductive load with a nearby capacitance brings the power factor of the area load closer to unity. This means a reduction in current, since now a smaller apparent power is needed to deliver the real power demanded by the load, which in turn causes a reduction in the voltage drop along the line according to Ohm's law.

Recall that a circulating current is required to serve a reactive load, which has to travel between wherever reactive power is "generated" and "consumed," effectively swapping stored energy between electric or magnetic fields during certain portions of the cycle. If reactive power is injected at a transmission-level generator bus to match the reactive demand, this circulating current must travel throughout the transmission and distribution system to reach any given load. If reactive power is injected locally instead, this circulating current does not need to accompany the real power on its way from the power plant. Besides saving line losses, this reduces the voltage drop along the distribution feeder. For this reason, distribution utilities may install capacitor banks near loads known to have a high inductive component.

Figure 7.19 illustrates the effect of a capacitor on the voltage profile. Rather than a sudden step up, it reduces the slope of voltage drop along the feeder.

The case sketched in Figure 7.19 is consistent with the intuition of reducing voltage drop by reducing total current. This interpretation is adequate for some practical purposes, but it does not capture the whole story. In fact, capacitors don't just mitigate voltage drop; they can also create an absolute voltage *rise* along the circuit, in a way that is not at all intuitive. To explain the voltage rise phenomenon, it is necessary to account for complex voltages and currents (i.e., magnitudes

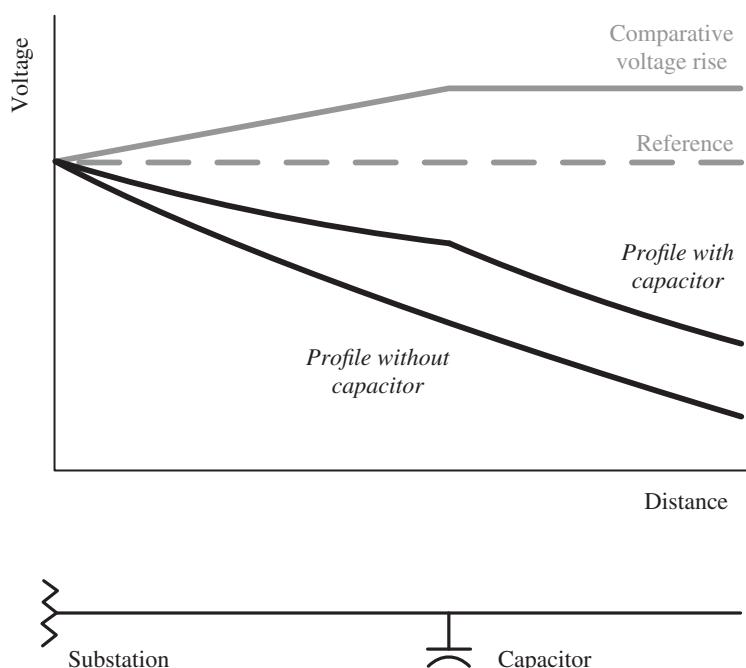


Figure 7.19 Effect of a capacitor on the voltage profile of a radial distribution feeder.

and phase angles) as well as the complex impedance of the line itself. We defer this important discussion to Chapter 9, where voltage drop is shown in phasor diagrams such as Figure 9.10. The *LinDistFlow* equations in Section 12.6 also help conceptualize and estimate voltage drop in relation to real and reactive power flows on radial distribution lines.

One fact to keep in mind is that a voltage rise across a transmission or distribution line is effected by capacitance if (and only if) the line itself is inductive in character. Since almost all lines in practice are overwhelmingly inductive, the conventional wisdom that “capacitors raise voltage” is broadly applicable.

7.5 Protection

7.5.1 Basics of Protection and Protective Devices

Circuit *protection* refers to a scheme for disconnecting sections or components of an electric circuit in the event of a *fault*. A fault means that an inadvertent electrical connection is made between an energized component and something at a different potential. If two conductors are touching directly, this makes a short circuit, or connection between two points that were initially at different potentials with essentially zero resistance in between.

The basic types of faults in power systems are *phase-to-ground* and *phase-to-phase* faults. A phase-to-ground fault means that one or more conductors make electrical contact with the ground, or point of 0-V potential, such as a line coming in contact with a tree (which, owing to its moisture, will conduct a current to ground). A phase-to-phase fault means that two different phases (or, rarely, all three) come into direct or indirect contact with each other, for example, if a bird with a large wingspan touches two conductors simultaneously.

When analyzing what would happen during any conceivable fault, the main quantity of interest is the *fault current*. The fault current is determined by the fault impedance (i.e. the electrical impedance of whatever object or creature is between the two points that are inadvertently connected) and by the *source impedance*, or ability of the power source to sustain the voltage even under an abnormally high current.

A fault is always something to be avoided, not only because it implies a wasteful flow of electric current, but because there is a risk of fire or electrocution when current flows where it was not intended to go. The object of circuit protection is to reliably *detect* a fault when it happens, and to interrupt the power flow to it, *clearing* the fault. Most commonly, a fault is detected by the magnitude of its associated current, though it may also be sensed, or its presence deduced, in other ways. For example, protective devices might look for a phase imbalance, some unusual voltage difference between circuit components, or an unusual ratio of voltage and current.

The simplest protective device that can detect an overcurrent and interrupt a circuit is the *fuse*. It basically consists of a thin wire that melts when the current is too high. Fuses come in a wide range of sizes—from tiny ones to protect small electronics, to those installed in homes in the early 20th century before circuit breakers became standard, to large *fused cutouts* on utility poles.

While fuses are very reliable, they have practical drawbacks. First, it takes some minimum amount of time before the wire heats up enough to melt. Once installed, it is not possible to change the sensitivity of a fuse, or how much current it will take to melt it. Then, once the wire has melted, it has to be physically replaced before the connection can be reestablished; it cannot be reset. In the utility setting, this means a “truck roll”—bringing a crew on location—that can

take hours to restore service. Fuses are used for radial feeders in distribution systems, generally for a lateral feeder where it connects to the main (see Section 7.1). In these situations, the desired sensitivity of the fuse is fixed, and the time delay for restoring service is considered acceptable because only a small number of customers are affected.

Circuit breakers differ from fuses in that they have movable contacts that can open or close the circuit. This means that a circuit breaker can be reset and reused after it opens. The mechanical opening or *tripping* of the breaker is prompted by a *relay*, which conveys a signal based on whether a measurement (such as the line current) is above some predetermined threshold value. Such a relay can have multiple settings, depending on the desired sensitivity.

While a circuit breaker can move more quickly than a melting fuse, it still takes a certain amount of time for a current to persist before the relay will actuate. For a standard overcurrent relay, this time is inversely related to the magnitude of the current. At the same setting, the relay could be tripped either by a very large current for a very short amount of time or by a smaller current for a longer duration. The sensitivity of relays and fuses is thus characterized by a *time–current* curve that indicates the combination of current and duration that will cause a trip.

A sample curve for a certain relay is shown in Figure 7.20. Note that both current and time are plotted on a logarithmic scale. For a large fault current, the *fault clearing time* should be a fraction of a second, on the order of several or tens of cycles. In some situations it may be problematic to distinguish between what is a fault current and what is just a high load current. This is especially true for *high-impedance faults*, where whatever is making the improper connection between a conductor and ground does not happen to conduct very well, and the fault current is therefore small.

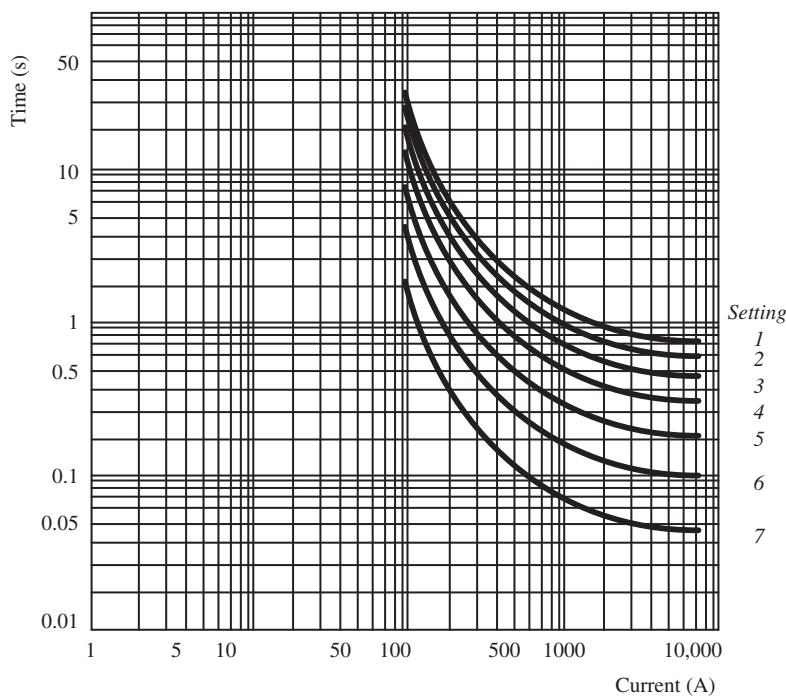


Figure 7.20 Sample time–current characteristic of a relay.

The problem of recognizing small fault currents is addressed by another method of fault detection that compares the currents on the two or three different phases, or between a phase and return flow. Even a small fault current from one of the phases to ground will result in a difference between the currents in each conductor. This difference is detected by a *differential relay*, which sends a signal to an actuator that opens the circuit. Differential relays are used in transmission and distribution systems, but also in the familiar *ground-fault circuit interrupters* (GFCIs) in residential settings. GFCIs are used where outlets are close to water sources, or where there is a danger of appliances coming in contact with water (which could cause persons who are also in contact with the water to be electrocuted); U.S. electrical code requires them in kitchens and bathrooms. Not only can GFCIs detect a smaller fault current than a conventional circuit breaker, they can also operate more quickly, within a few cycles—a fraction of a second that could make a vital difference in some situations.

Switches and circuit breakers in power transmission and distribution systems are collectively referred to as *switchgear*; they serve the purposes of deliberately isolating individual pieces of equipment (say, for maintenance) and to automatically isolate portions of the system (in case of a fault). An important distinction between a circuit breaker and a regular switch is that the breaker can safely interrupt a fault current, which may be much larger than a normal load current. In order to do so, circuit breakers must be specifically designed to control and extinguish the arc of *plasma* (see Section 1.1.4) drawn as the contacts separate.

For simple air switches, the design crux lies in the shape of the contacts. To be able to operate at higher voltages and currents, circuit breaker contacts can be immersed in a tank of nonconducting fluid such as transformer oil or sulfur hexafluoride (SF_6) that is also difficult to ionize and rapidly quenches the arc.²² More effective yet, the contacts can be placed in a vacuum. Finally, there are *air-blast circuit breakers* and *puffer-type arc interrupters* in which a burst of compressed gas is precisely directed at the arc to quench it (in a sense, mechanically, by pushing ionized molecules away from each other).

In the operation of a circuit breaker, time is of the essence—both in terms of when the breaker first actuates, and in terms of the physical movement of the contacts. The key problem is that the ionization of the medium between the contacts, which forms and sustains the arc, depends on both the voltage across the gap and its width.²³ When the metallic contacts initially move apart, they will necessarily draw an arc between them—just like the arc that can be seen when an operating appliance is unplugged from its outlet, only bigger. But within half a cycle, the alternating voltage and current become zero, and the arc will naturally extinguish. (Note that this characteristic is peculiar to a.c. and explains why high-voltage d.c. circuit breakers are more difficult to engineer.) By the time the voltage rises again in the opposite direction—and we are talking about mere milliseconds—the breaker contacts may or may not have already moved far enough apart to prevent the formation of a new arc, or *restrike*. Restriking is undesirable because it wears out the breaker contacts, in addition to prolonging the time before the fault is cleared. Ideally, the contacts can be physically moved fast enough, so that the growing distance between them outpaces the sinusoidally increasing voltage. The properties of the circuit are relevant here, too, because any reactance will affect the relative timing between current and voltage.

²² The fluid should consist of highly symmetrical molecules without free electrons that do not readily break apart under the influence of an electric field. Quite unrelated to its excellent insulating characteristics, SF_6 happens to be a powerful greenhouse gas; see also Section 8.2.

²³ For example, the ionization potential of air is on the order of a million volts per meter, so a voltage difference of tens of kilovolts forms an arc across centimeters.

In addition to the above complexities, consider that arc plasma temperatures are on the order of tens of thousands of degrees Celsius; that large mechanical components are accelerated to high speeds within thousandths of a second; and that pressurized quenching gas flows at supersonic speeds. This should help us appreciate that the design of power switchgear is a serious business indeed.

Many times faults are *transient*, meaning that their cause disappears. For example, lightning strike may cause a fault current that will cease once the lightning is over; power lines may make contact momentarily in the wind; or a large bird may electrocute itself across two phases and drop to the ground, removing the connection. In these situations, it is desirable for the circuit to be restored to normal operation immediately after the fault disappears. For this purpose, *reclosing* breakers (or *reclosers* for short) are used. The idea is that the breaker opens when the fault is detected, but then, after some time has passed—the *reclosing time*—closes again to see if the fault is still there. If the current is back to normal, the breaker stays closed and everything is fine; customers have only suffered a very brief interruption. This scenario is illustrated in Figure 7.21.

If the fault current is still there, the recloser opens again. This cycle may repeat another time or two, and if the fault persists on the last reclosing attempt, the breaker stays in the open or *lockout* position until it is reset. This process is illustrated in Figure 7.22, where the recloser makes only one reclosing attempt. When the fault current is initially detected, the recloser opens after just a few cycles. (Note that the drawing is not to scale, as the reclosing interval could last for many more cycles.)

The reclosing time and number of attempts can be adjusted as appropriate. In distribution systems, reclosing times tend to be much longer than in transmission systems—five seconds, perhaps, as compared to half a second. Primarily, this is because distribution equipment tends to be closer to the ground and more exposed to environmental factors that take a little longer to go away, including unfortunate incidents with animals. Another consideration is the number of customers exposed to the interruption, which is of course much greater for a transmission fault. This shifts the desired reclosing time in transmission systems toward the short side, especially if it can be made brief enough to escape customers' notice completely. Like time–current settings on a breaker, the choice of recloser settings illustrates the fact that any circuit protection inherently involves some trade-off between safety and convenience.

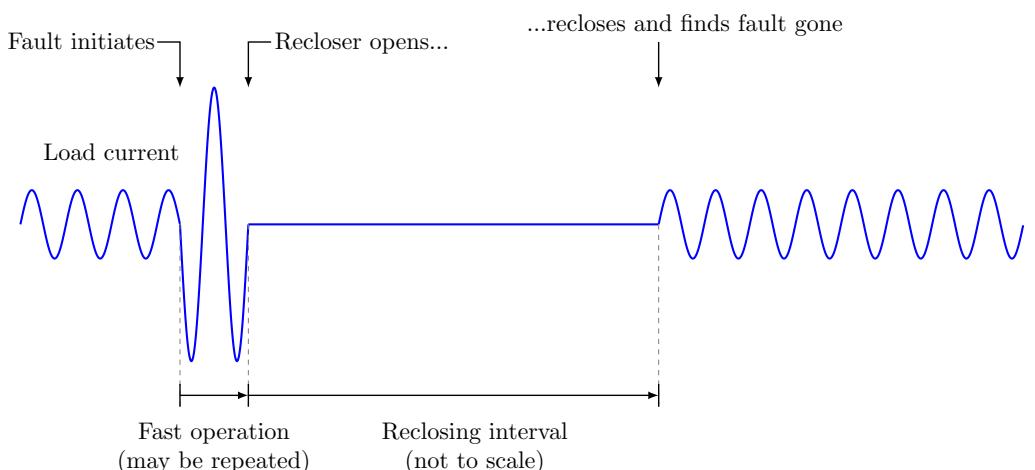


Figure 7.21 Recloser operation with transient fault.

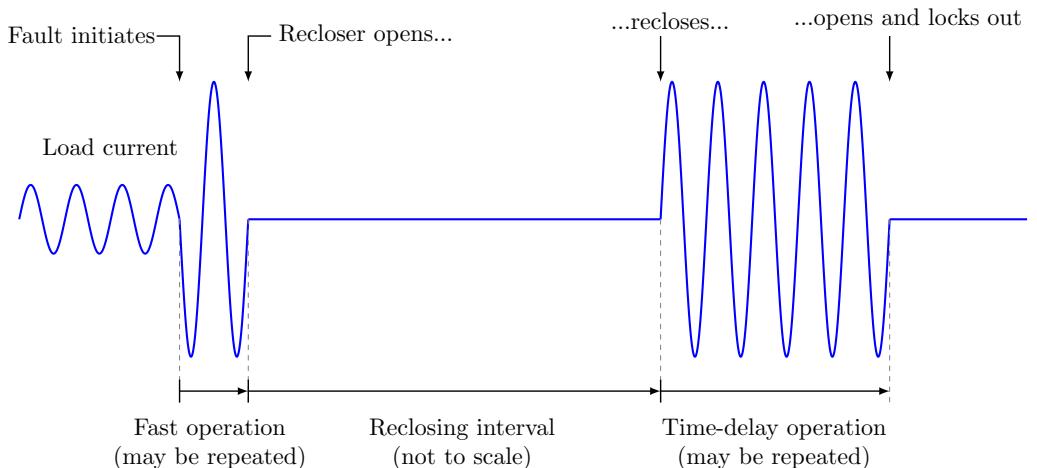


Figure 7.22 Recloser operation with a permanent fault. There could be several additional reclosing attempts (fast or slow) before locking out.

7.5.2 Protection Coordination

In order to minimize service interruptions, power system protection is carefully designed to interrupt the circuit as close as possible to the fault location. There is also redundancy in protection, meaning that in the event one breaker fails to actuate, another one will. With both of these considerations in mind, protection throughout the system is coordinated so that for any given fault, the nearest breaker will trip first. Such a scheme is analyzed in terms of *protection zones*, or sections of the system that a given device is “responsible” for isolating. These zones are nested inside each other, as illustrated in Figures 7.23–7.26.

In such a scheme, any one protective device may simultaneously serve as the primary protection for its own zone and as backup for another. For example, Fuse 1 is the primary protection for the section of line between it and Fuse 2, while also serving as a backup in case Fuse 2 should fail to clear a fault in its own protection zone. Of course, we do not wish for Fuse 1 to melt and unnecessarily

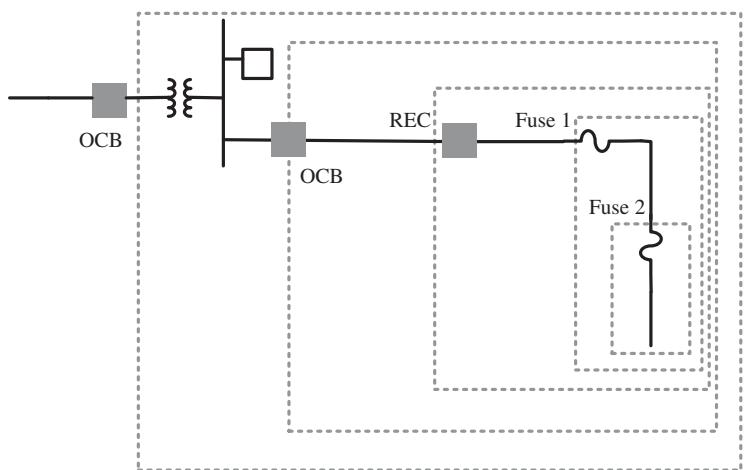


Figure 7.23 Example of protection zones with oil circuit breakers (OCB) and recloser (REC).

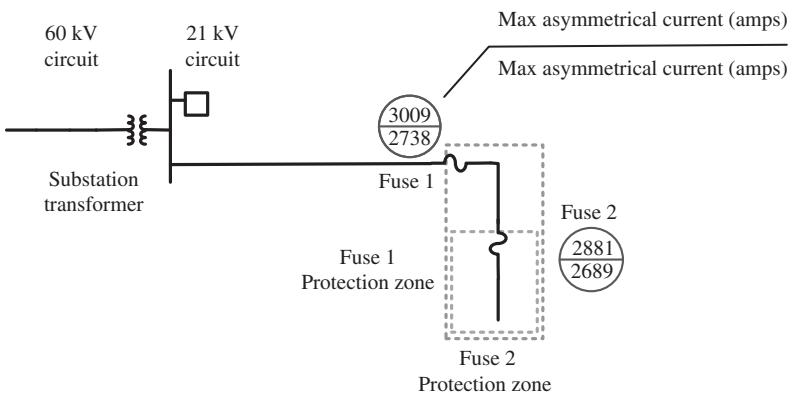


Figure 7.24 Protection zones—fuses.

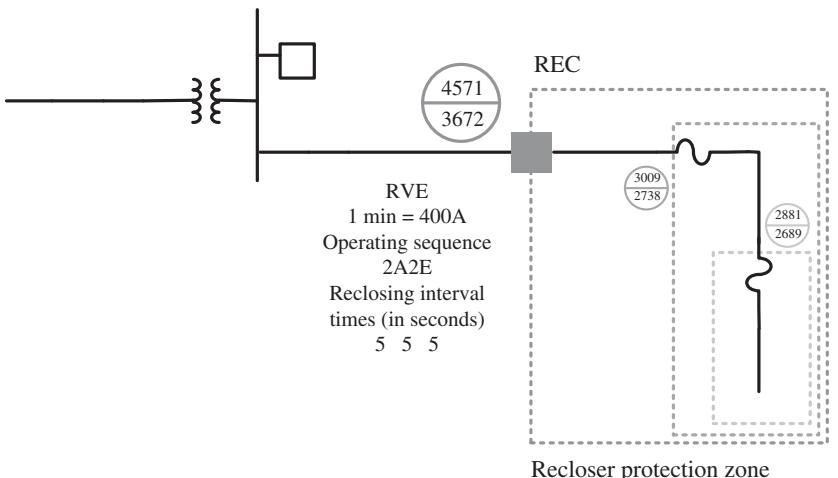


Figure 7.25 Protection zones—reclosers.

inconvenience customers in its own protection zone as the result of a problem beyond Fuse 2 on the circuit. For this reason, Fuse 1 ought to be less sensitive, that is, tolerate a greater current than Fuse 2.

The illustrations in Figures 7.23–7.26 show that the maximum current is greater for devices upstream in the circuit, which is also necessary simply because there is more load connected whose normal current must flow through them. The terms *symmetrical* and *asymmetrical* are explained in Section 7.5.3 below.

The sample specifications on the recloser and circuit breaker relay highlight the crucial time dimension of the protection problem: it is important not only *whether* a device operates, but precisely *when*. The coordination among protective devices is analyzed graphically using time–current curves. For example, Figure 7.27 illustrates coordination curves for a fuse and a recloser. Here we introduce two additional subtleties. First, it is not desirable for a fuse to melt partially, and then be left in an unknown, damaged condition. The time it takes for the fuse to melt completely is indicated by the vertical separation between the *minimum melting time* and the *maximum clearing*

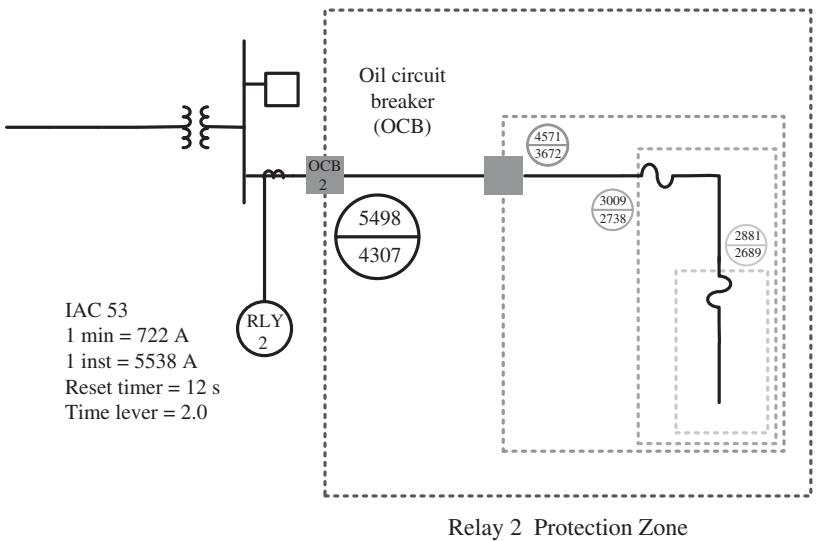


Figure 7.26 Protection zones—relay.

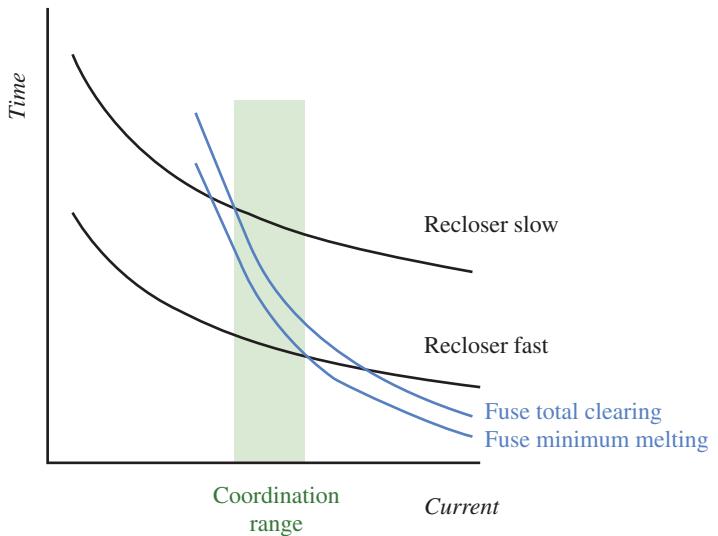


Figure 7.27 Sample coordination with time–current curves for a fuse and a recloser.

time. Second, the recloser's initial (fast) and repeated (slow) action are indicated. The range of fault currents where the two devices are properly coordinated lies in the shaded area.

Here's how we would expect events in Figure 7.27 to play out: A fault occurs, and the recloser interrupts it by opening according to the "fast" curve. Some time is allowed to elapse (say, several seconds—not shown on the diagram) during which power is interrupted, and during which the cause of the fault might disappear naturally. The recloser re-energizes the circuit (not shown). If the fault has gone away, all is well and service is restored. Because the recloser actuated quickly, there is not yet any damage to the fuse, even at the maximum expected fault current. If the fault is still present, the current returns. For repeated action, the recloser is set to operate on the "slow"

curve. The purpose is to now delegate the fault-clearing job to the fuse, assuming that the fault is downstream from the fuse. Here we want to ensure that the fuse has time to melt completely, even at the minimum expected fault current. If so, the fault is now cleared by the fuse, with a minimum number of customers interrupted. If the fault lies somewhere between the fuse and the recloser, the fault current does not travel through the fuse (assuming we are in a radial system where the substation is the only source!). In this case, the fuse does nothing and the recloser trips again, isolating the fault and (unavoidably) interrupting more customers until crews arrive to address the problem. The curve intersections determine the minimum and maximum currents for which the devices are coordinated.

The main take-away is that protection coordination involves multiple variables to be considered in a scheme that will perform safely and reliably, yet without causing nuisance interruptions, under a range of foreseeable circumstances. The above illustration also highlights the significance of radial distribution systems with unidirectional power flow, and why utilities are so keen to ensure that customer-owned generation disconnects when there is a fault on the system (see Section 15.2.2).

In a network, protection coordination becomes even more challenging, because here the roles of primary and backup protection (i.e., which one trips first) must be reversed depending on what side the fault is on. Yet the only means of discriminating the distance to a fault is by the impedance of the line in between, which is not easily estimated in real time.²⁴ This complexity alone is sufficient reason for the majority of power distribution systems to be laid out radially. It also explains why protection engineering is a task for highly specialized experts who draw not only on formal analysis but also on experience and intuition to make it work in practice.

7.5.3 Unsymmetrical and Asymmetrical

Faults can be *symmetrical* or *unsymmetrical*, depending on whether they are the same or different on each of the three phases. The vast majority of faults are unsymmetrical, since it is much more likely that some unintended electrical contact involves just one or two phase conductors, instead of all three equally.

Because there is no longer three-phase balance, it is necessary to decompose fault currents into *symmetrical components* for analysis (see Section 4.2). Here we identify positive-, negative-, and zero-sequence impedances of the devices in a faulted circuit, and on this basis can identify the positive-, negative-, and zero-sequence currents. This allows us to analyze an extremely unbalanced current as the mathematical superposition of three balanced three-phase currents.

In a particularly nonobvious usage unique to power engineering, the term *asymmetrical* means something completely different than *unsymmetrical*. While unsymmetrical refers to the three phases, asymmetrical refers to the evolution in time of the alternating fault current. Specifically, it describes the extent to which a fault current initially predominates in one direction over the other, rather than symmetrically alternating in the positive and negative direction like regular alternating current.

To understand why there would be asymmetry, we need to look at the *transient* behavior of a faulted circuit explicitly in the time domain, including the very short time scale of less than a cycle, rather than the steady-state behavior that we analyze with phasors. Inductance will play a special role here.

When there is a fault, most of the impedance in the circuit will come from the conductors and especially the transformers. As discussed in Section 8.5, these are typically characterized by an

²⁴ Some of the advanced measurement technologies discussed in Section 16.2 can be applied for this purpose.

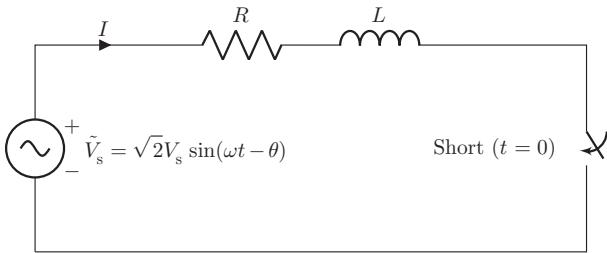


Figure 7.28 $R-L$ circuit of fault loop shorted at $t = 0$.

inductive reactance and have very little resistance. By essentially short-circuiting the regular load, the beginning or *inception* of a fault is equivalent to suddenly eliminating the resistance in a circuit, leaving only the inductive reactance of the transmission and distribution system. This situation is represented in Figure 7.28, where the inception of the fault is equivalent to closing the switch, causing a sudden increase in current.

Recall that the current flowing through an inductor is limited by its rate of change, which is proportional to the source voltage:

$$v_L(t) = L \frac{di}{dt}$$

We may assume here that the grid voltage feeding the fault is essentially the same as before, even though in reality it will be diminished somewhat due to source impedance (which causes the voltage sags that usually accompany faults). As the fault starts, the resistance suddenly drops away, and the inductor is instantly exposed to a much greater voltage across it.

What matters now, as we analyze the current driven by a sinusoidal voltage in the time domain, is the precise timing of the fault's *inception* within the cycle, illustrated in Figure 7.29. As long as there is a voltage in the same direction, the inductor current will continually increase.

Normally, in the steady-state condition with symmetrical a.c., the voltage and current each spend exactly half as much time in the positive and negative direction. The current is at its maximum negative value when the voltage crosses zero. It increases throughout the positive voltage cycle (180°), but spends the first half of that cycle (90°) just getting back to zero. The current then grows in the forward direction and reaches its maximum just when the voltage crosses zero again. Formally, the current magnitude at time T can be expressed as the integral of the above equation:

$$i(T) = \frac{1}{L} \int_0^T v_L(t) dt$$

Consider now what happens if, perchance, the fault began at the instant that the voltage crossed zero. The fault current will build up throughout the entire half-cycle of 180° . Only when the voltage changes direction will the current reach its maximum and start to decrease. But this yields a current that is twice as big as the normal amplitude of a symmetrical current (Figures 7.28 and 7.29)!

After the voltage reverses, the current will diminish, but it will take the entire next 180° of negative voltage just to get back to zero. What we have, then, is an alternating current with the peak-to-peak amplitude as expected from the value of the inductance and frequency, but shifted upward to sit right above the x -axis. This is described as a *d.c. offset*, because it is mathematically equivalent to adding a constant d.c. component to the symmetrical a.c. Due to the damping effect of the series resistance, this condition is not sustained, and the actual current decays to its symmetrical a.c. value after some number of cycles, as illustrated in Figure 7.30. But the initial peak

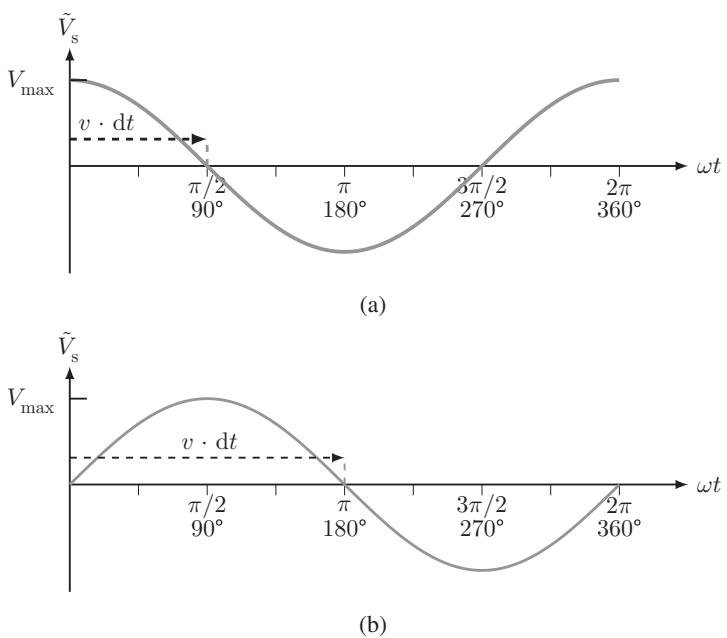


Figure 7.29 The role of fault inception: current increases during the entire time that voltage remains positive on the first cycle. (a) Fault incepts at $\theta = 90^\circ$ when $v_s = V_{max}$. (b) Fault incepts at $\theta = 0$ when $\tilde{V}_s = 0$ (worst case).

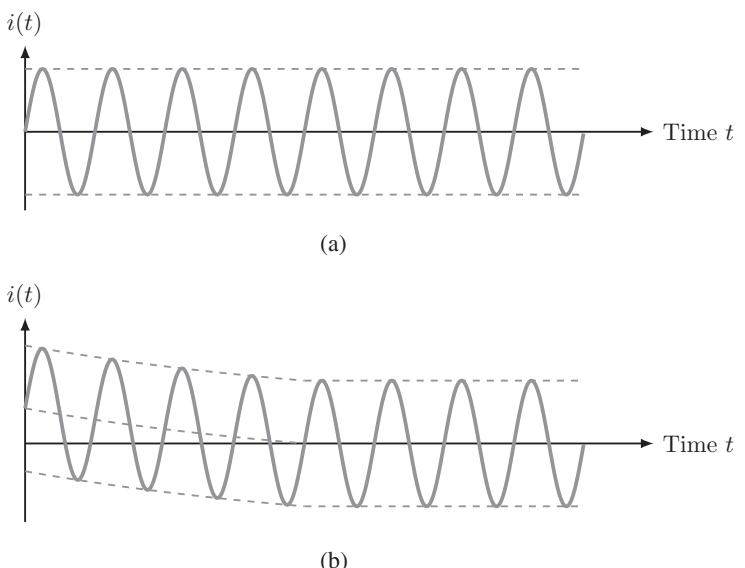


Figure 7.30 Symmetrical (a) and asymmetrical (b) current with a decaying d.c. component due to the timing of the fault's inception.

remains a practical concern, in large part because the instantaneous current is associated with a magnetic field and thus a mechanical force on the switchgear.

The fault inception at the 0° voltage angle is a worst-case scenario, where the initial peak current exceeds the regular, symmetrical amplitude by a factor of 2. If the fault happens to start at the maximum voltage at 90° , there is no d.c. offset at all. For the inception times in between, or between 90° and 180° , there is some partial amount of offset and increase in the amplitude, and values between 180° and 360° work the same way except with the offset in the opposite direction.

The actual peak also depends on the ratio of resistance and inductance in the faulted circuit. A *peak factor* K is defined as the ratio of the first asymmetrical current peak to the symmetrical rms value of the current. Since this definition includes a factor of $\sqrt{2}$ for the ratio of an amplitude to an rms value, the worst-case value of K is $2 \times \sqrt{2} \approx 2.8$.

It is also worth noting that in a three-phase system, if the fault inception is exactly simultaneous for all three phases, it will necessarily hit each phase at a different point in its cycle. In any case, the simple qualitative take-away is that fault currents can be treacherous.

The phenomenon of potentially doubling an instantaneous current seems important. Why haven't we learned about this before—doesn't this happen every time you start up an a.c. circuit with inductance? Yes, but when a transmission line or transformer is being energized, we know what to expect; the equipment is designed to handle the current, and components are connected sequentially with great care. Note that the d.c. offset decays quickly over time, so it will not cause a thermal issue. Any resistance at all in the circuit will have the effect of driving the current peaks down, and will stabilize the alternating current into a symmetrical pattern. Therefore, analysis of transient asymmetrical currents is not necessary in the normal operational context.

Protective devices, by contrast, must be designed with multiple criteria in mind: they must tolerate the heat that develops as a fault current is sustained over time, and they also must withstand the mechanical forces associated with instantaneous peak currents, which can do damage despite their very short duration.

Derivation

Some readers may be familiar with the equations for an "RL" circuit with direct current. The voltage v_L across the inductor, in series with a resistor and supplied by a d.c. voltage v , is given by the exponential

$$v_L(t) = v e^{-\frac{tR}{L}}$$

where

$$v = v_L(t) + v_R(t)$$

and the current through the series combination is

$$i(t) = \frac{v}{R} \left(1 - e^{-\frac{tR}{L}} \right)$$

where the ratio $\tau = L/R$ is called the time constant (note that the exponent must be dimensionless). At time $t = 0$, the exponential term is 1, and the entire voltage is observed across the inductor; since there is no current yet, the resistor does not sustain any voltage drop. As t proceeds, the current through the series combination builds up, as does the voltage drop across the resistor. Eventually, the exponential term vanishes, the resistor is seeing the entire voltage drop, and the inductor does nothing but conduct whatever amount of current the resistor allows.

Now consider that if we zoom in close enough on our time scale, one a.c. cycle seems to last for an eternity, and any portion of it basically appears as a d.c. situation. The alternating characteristic

comes into play when we pay attention for more than one cycle. Thus, in analyzing a transient condition, the fast part appears as d.c. and the slow part as a.c. Therefore, it makes some intuitive sense that the decaying exponential associated with the instantaneous start-up of the RL circuit should be somehow combined with the a.c. waveform that describes the steady state.

The rigorous mathematical derivation is not trivial. We need to write a differential equation for current, given a sinusoidal voltage with angular frequency ω . Let's label the moment of fault inception with the voltage angle θ . Then we write

$$v(t) = \sqrt{2}V_{\text{rms}} \sin(\omega t + \theta) = Ri + L \frac{di}{dt}$$

where we have added the voltage drop $v_R = Ri$ to the previous v_L . Since any load current present at the beginning is much smaller than the fault current, we set initial conditions to $i(0) = 0$. Also, $v(0) = \sqrt{2}V_{\text{rms}} \sin \theta$. It will be convenient to define the angle by which current lags voltage, $\theta_Z = \tan^{-1}(X/R)$, where the subscript Z refers to the impedance in the circuit.

The solution to this differential equation, which explicitly describes current as a function of time, is

$$i(t) = \sqrt{2}I_{\text{rms}} [\sin(\omega t + \theta - \theta_Z) - \sin(\theta - \theta_Z)e^{-\omega t R/X}]$$

where the exponent has been formatted in terms of more practical quantities using ω .

In the worst case where $\theta = 0$ and the fault impedance is purely inductive, $\theta_Z \approx 90^\circ$ and we can rearrange terms to write

$$i(t) = \sqrt{2}I_{\text{rms}} [e^{-\omega t R/X} - \cos \omega t]$$

This function will peak after half a cycle, when $\cos \omega t = -1$, at which point the exponential term will not yet have decayed very much.

Problems and Questions

- 7.1 Explain the terms *load factor* and *load density* in your own words. How has each affected the design of electric power systems?

- 7.2 A three-phase load of 100 kVA with power factor 0.8 lagging is supplied by conductors with $R = 0.0525 \Omega/1000 \text{ ft}$ and $X = 0.031 \Omega/1000 \text{ ft}$. The voltage seen by the load at the receiving end of the line is $V_R = 280 \text{ V}$ (line-to-neutral).
 - (a) What is the voltage drop over a distance of 1000 ft, to two significant figures?
 - (b) What are the line losses, in kW and percent of power delivered, to one significant figure? Comment.
 - (c) If the losses are not to exceed 1%, what is the maximum conductor length, to one significant figure?
 - (d) Suppose the same conductor is used with a higher source voltage, such that the new receiving end voltage supplying the 100-kVA load is $V_R = 2.7 \text{ kV}$. What is the voltage drop over 1000 ft for the 2.7-kV case?
 - (e) Repeat line losses and maximum conductor length for the 2.7-kV case. Comment on the practical relevance of your result.

- 7.3 Suppose a transmission line is 100 km long and uses an aluminum conductor material with resistivity $\rho = 2.7 \times 10^{-8} \Omega \cdot \text{m}$. Your design objective is that thermal losses should

not exceed 100 kW when the line is delivering 100 A per phase. Calculate the required conductor diameter.

- 7.4** Given the density of aluminum is 2.7 g/cm³, calculate the weight of a 100-m length of conductor for the previous problem.
- 7.5** Look up resistivity and density values for copper and silver. Let's pretend that money is no object (and neither is tensile strength). Would either of these metals allow the use of a lighter weight transmission conductor than aluminum, for the same amount of losses? Does the answer to this question depend on conductor size?
- 7.6** Let's think about the mechanics of transmission lines a bit more. Consider a pure aluminum conductor with 2000 kcmil cross-section.
- What is the mass in kg of a 200-m long span of this conductor?
 - Suppose that at the point where this conductor is suspended from the insulator at the transmission tower, it makes a 15° angle with the horizontal. What is the tension in the conductor, in newtons? (Hint: Weight is the vertical component of the tension vector.)
 - What other forces besides gravity should the design engineer be thinking about?
 - Now suppose that you wish to build this transmission line across a body of water, where it is very expensive to construct concrete piers, so you want to space the towers farther apart. However, you also need to maintain the same ground clearance (i.e., height of the lowest point of the conductor in the middle of the span). What are your design options?
 - What role does current on the transmission line play for the above calculation?
- 7.7** Referring to Table 7.1, confirm that the surge impedance loading is consistent with the other data given for each of the three sample transmission lines.
- 7.8** Qualitatively, list the considerations for sizing a fuse or selecting a time-current trip setting of a circuit breaker relay. What questions would you ask about the circuit on which these are installed? What could happen if you err on either side (too small or too large a rating)?
- 7.9** Refer to the time-current characteristic in Figure 7.20.
- Could this relay be used in a situation where the rated load current is 100 A? Explain.
 - On Setting 5, how long would it take the relay to trip for a fault current of 500 A?
 - What is the fastest trip time, in cycles at 60 Hz, for this relay?
- 7.10** Refer to Figures 7.23–7.26. Suppose a fault occurs somewhere between Fuse 1 and Fuse 2. Explain how you expect each device to act: Fuse 1, Fuse 2, the recloser, and the oil circuit breaker.

8**Transformers****8.1 General Properties**

A transformer is a device for changing the voltage in an a.c. circuit. It basically consists of two conductor coils that are connected not electrically but through magnetic flux. As a result of electromagnetic induction, an alternating current in one coil will set up an alternating current in the other. However, the comparative magnitude of the current and voltage on each side will differ according to the geometry, that is, the number of turns or loops in each coil.

Consider the diagram in Figure 8.1 of a very simple transformer. The coil on the left-hand side, which we will label as the *primary* side, might be connected to a power source such as a generator, while the right-hand or *secondary* coil would supply a load. Looking at the electrical connections, we see two separate circuits in this diagram: the circuit between the generator and the transformer (with current I_1 and voltage V_1) and the circuit between the transformer and the load (with current I_2 and voltage V_2). Yet electrical power is somehow transmitted across the transformer from the generator to the load. Although a small percentage of the power will be dissipated as heat inside the transformer, almost all the power is transferred from one winding to another through the magnetic core. Essentially, the same amount of power goes into the transformer as goes out.

Recalling that electrical power is given by the product of current and voltage, we can state that the product of current and voltage on each side of the transformer must be the same. Referring to the labels in the diagram, we would write

$$I_1 \cdot V_1 = I_2 \cdot V_2$$

However, based on examining only the electrical connections, we cannot say anything more about how the voltage and current relate. For this, we must look at what happens with the magnetic flux and induction inside the transformer.

First, in the primary coil, the a.c. supplied by a generator produces a magnetic field, or flux, inside the core of the coil. Like the conductors inside a generator, transformer coils are wound around a core of magnetically susceptible material, generally some type of iron or electrical steel, to channel and enhance the magnetic field.¹ Real transformers are more efficient if the core forms a closed ring, unlike the U-shape in Figure 8.1, but the key principle is that the windings are linked by magnetic flux, which is guided by the core. The magnetic flux resulting from the current is proportional to both the current's magnitude and the number of turns in the coil.

Using the terminology introduced in Section 2.6, we can state that the product of the current I and the number of turns n in the winding gives the *magnetomotive force (mmf)*, which produces

¹ The choice of material for a transformer core is very important in practice for performance and efficiency, but the basic principle of a transformer works even with just air in the core.

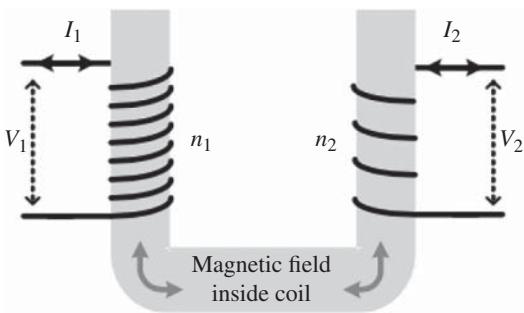


Figure 8.1 Transformer concept. A realistic transformer performs better if the core forms a closed loop, as shown in Figure 2.13.

the *magnetic flux* (denoted by Φ , Greek capital phi) inside its core. The flux is enhanced because the iron core has a high magnetic permeability, or a low magnetic *reluctance* (denoted by script \mathcal{R}). In equation form,

$$\text{mmf} = nI \quad (8.1)$$

and

$$\Phi = \frac{\text{mmf}}{\mathcal{R}} \quad (8.2)$$

Because the reluctance of the transformer core is significantly lower than that of the surrounding air, the flux is in effect “captured” inside the core, meaning that the leakage flux outside the core is small.

As the current on the primary transformer winding alternates, the flux inside the core changes direction back and forth 60 times per second (for 60-Hz a.c.). But this same changing flux links the secondary winding around the same core. Proportional to the rate of change of this flux, an electromotive force (emf) is induced in the secondary winding. A current then flows in this winding as determined by the impedance connected to it. The emf is also proportional to the number of turns, denoted by n , of this winding around the core. This makes sense in that an emf is induced in each turn, and the successive turns of the same conductor are effectively in series, so that the voltages or emfs are additive.

With regard to changing the voltage from the primary to the secondary side, the key measure is the *turns ratio*, or ratio of the number of turns in the primary and secondary winding. If both primary and secondary windings had the same number of turns, the voltage would be the same on either side of the transformer. In order to increase or step up the voltage, the number of turns must be greater on the secondary side; conversely, to step down voltage, the number of turns must be less. For example, the primary winding in Figure 8.1 is shown as having eight turns and the secondary winding four, so that the voltage would be stepped down by a factor of two. A real transformer would have hundreds of turns in each winding so as to make the magnetic induction process more efficient, but as long as the ratio of primary to secondary turns remains the same, the effect on voltage is the same. Because power is conserved (to a first approximation), and since power equals voltage times current, the currents through either winding are inversely proportional to the voltage and the number of turns. In equation form,

$$\frac{V_1}{V_2} = \frac{n_1}{n_2} = \frac{I_2}{I_1} \quad (8.3)$$

The transformer in this example would be called a *step-down* transformer if the greater number of turns is on the primary side, which is defined as the side to which the power source is connected.

The identical device, if connected to a generator on the right and a load on the left, would be called a *step-up* transformer.

The voltage on the secondary side can be deliberately changed if there is a movable connection between the winding and the conductor. Such a connection is called a *transformer tap*. Depending on where the conductor taps the secondary winding, this circuit will effectively “see” a different number of turns, and the transformer will have a different effective turns ratio. By moving the tap up or down along the winding, the voltage can be adjusted. Distribution transformers, especially at the substation level, generally have *load tap changers* (LTCs) to adjust the connection in a number of steps. These LTCs are moved to different settings in order to compensate for the changes in voltage level that are associated with changes in load (see Section 7.4).

8.2 Transformer Heating

In a real transformer, some power is dissipated in the form of heat. A portion of these power losses occur in the conductor windings due to electrical resistance and are referred to as *copper losses*. However, so-called *iron losses* from the transformer core are also important. The latter result from the rapid change of direction of the magnetic field, which means that the microscopic iron particles must continually realign themselves—technically, their *magnetic moment*—in the direction of the field (or flux). Just as with the flow of charge, this realignment encounters friction on the microscopic level and therefore dissipates energy, which becomes tangible as heating of the material. Section 8.5 discusses these losses in more detail. The type and quality of steel used in transformer cores has a large effect on the amount of heat produced, and *grain-oriented electrical steel* is essential for high-power applications.

Taking account of both iron and copper losses, the efficiency (or ratio of electrical power out to electrical power in) of real transformers can be in the high 90% range. Still, even a small percentage of losses in a large transformer corresponds to a significant amount of heat that must be dealt with. In the case of small transformers inside typical household adaptors for low-voltage d.c. appliances, we know that they are warm to the touch. Yet they transfer such small quantities of power that the heat is easily dissipated into the ambient air (bothering only conservation-minded analysts, who note the energy waste that could be avoided by unplugging all these adaptors when not in use). By contrast, suppose a 10-MVA transformer at a distribution substation operates at an efficiency of 99%. A 1% loss here corresponds to a staggering 100 kW.

In general, smaller transformers like those on distribution poles are passively cooled by simply radiating heat away to their surroundings, sometimes assisted by radiator vanes that maximize the available surface area for removing the heat. Large transformers like those at substations or power plants require the heat to be removed from the core and windings by active cooling, generally through circulating oil that simultaneously functions as an electrical insulator.

The capacity limit of a transformer is dictated by the rate of heat dissipation. Thus, as is true for power lines, the ability to load a transformer depends in part on ambient conditions including temperature, wind, and rain. For example, if a transformer appears to be reaching its thermal limit on a hot day, one way to salvage the situation is to hose down its exterior with cold water—a procedure that is not “by the book,” but has been reported to work in emergencies. When transformers are operated near their capacity limit, the key variable to monitor is the internal or oil temperature. This task is complicated by the problem that the temperature may not be uniform throughout the inside of the transformer, and damage can be done by just a local hot spot. Under extreme heat, the oil can break down, sustain an electric arc, or even burn, and a transformer may explode.

A cooling and insulating fluid for transformers has to meet criteria similar to those for other high-voltage equipment, such as circuit breakers and capacitors: it must conduct heat but not electricity; it must not be chemically reactive; and it must not be easily ionized, which would allow arcs to form. Mineral oil meets these criteria fairly well, since the long, nonpolar molecules do not readily break apart under an electric field.

Another class of compounds that performs very well and has been in widespread use for transformers and other equipment is polychlorinated biphenyls, commonly known as PCBs. Because PCBs and the dioxins that contaminate them were found to be carcinogenic and ecologically toxic and persistent, they are no longer manufactured in the United States; the installation of new PCB-containing utility equipment has been banned since 1977.² However, much of the extant hardware predates this phaseout and is therefore subject to careful maintenance and disposal procedures, somewhat analogous to asbestos in buildings.

Introduced in the 1960s, sulfur hexafluoride (SF_6) is another very effective arc-extinguishing fluid for high-voltage equipment. SF_6 has the advantage of being nontoxic as well as chemically inert, and it has a superior ability to withstand electric fields without ionizing. While the size of transformers and capacitors is constrained by other factors, circuit breakers can be made much smaller with SF_6 than traditional oil-filled breakers. However, it turns out that SF_6 absorbs thermal infrared radiation and acts as an extremely potent greenhouse gas when it escapes into the atmosphere, and its concentration has increased substantially over the past decades.³ This surprising and unfortunate characteristic may motivate increasing restriction of SF_6 use.

8.3 Delta and Wye Transformers

As described in Section 8.1, each side of a transformer has one winding, or a single conductor with two ends to connect to a circuit. What happened to the three phases? Transforming three-phase power actually requires three transformers, one for each phase. These three may be enclosed in a single casing, labeled to contain a single three-phase transformer, or they may be three separate units standing next to each other, called a *transformer bank*. In either case, the three transformers are magnetically separate, meaning that they should not share magnetic flux among their cores, because the magnetic field or flux in each one oscillates in a different phase (i.e., with different timing). Figure 8.2 shows a large, three-phase transformer at a distribution substation.

The two distinct ways of connecting a set of single-phase loads to a three-phase system, the delta and the wye connections (see Section 6.2), also apply to connecting a set of transformers. In the wye connection, each transformer winding connects between one phase and ground, or a common neutral point shared by all three windings. In the delta connection, each winding connects between one phase and another. However, because the primary and secondary side of a transformer are electrically separate, the type of connection on either sides need not be the same. This provides for four distinct possibilities for a three-phase transformer connection: Δ - Δ , Y - Y , Δ - Y , and Y - Δ , all of which are commonly used. One of these four possibilities, the Δ - Y connection, is illustrated in Figure 8.3, showing both the shorthand symbolic representation and a schematic of the actual wiring. The letter n in the diagram stands for neutral.

The choice of delta or wye connection has implications for grounding and for voltage. For the Δ - Δ and Y - Y connections, the change in voltage from the primary to the secondary side is simply

² See <https://www.epa.gov/pcbs> (accessed February 2024).

³ See U.S. National Oceanic and Atmospheric Administration (NOAA), “Trends in Atmospheric Sulfur Hexafluoride,” https://gml.noaa.gov/ccgg/trends_sf6/ (accessed February 2024).

Figure 8.2 Transformer at a distribution substation. Source: Courtesy of Pacific Gas & Electric.



given by the turns ratio. If the connections on the primary and secondary sides differ, though, the voltage is modified, because the transformer is in effect converting a phase-to-phase to a phase-to-neutral voltage, or *vice versa*. Thus, a Δ -Y or Y- Δ connection introduces an increase or decrease of $\sqrt{3}$ to be factored in along with the turns ratio.

Suppose the delta-wye transformer in Figure 8.3 is connected on its left-hand (primary) side to a transmission line with 115-kV phase-to-phase and 66.4-kV phase-to-neutral voltage. This transformer has a turns ratio of 1:1, so that the voltage across any pair of windings is identical. On the primary side with the delta connection, this voltage corresponds to the phase-to-phase voltage, 115 kV. This same voltage on the secondary side with the wye connection, however, is now made to correspond to the phase-to-neutral voltage. Because the relationship between phase-to-phase and phase-to-ground voltages is mathematically fixed, the phase-to-phase voltage on the secondary side will be raised by a factor of $\sqrt{3}$, to 199 kV. Comparing the phase-to-phase voltages on the primary and secondary side, we find that the delta-wye transformer has effectively increased voltage by a factor of $\sqrt{3}$. A wye-delta transformer has the opposite effect.

Example

A wye-delta transformer with a 10:1 turns ratio steps voltage down from a 230-kV transmission line to a distribution circuit. What is the voltage on the secondary side?

Transmission-line voltages are labeled on the basis of phase-to-phase measurements. The wye connection on the primary side indicates that on its primary side, the transformer “sees” the

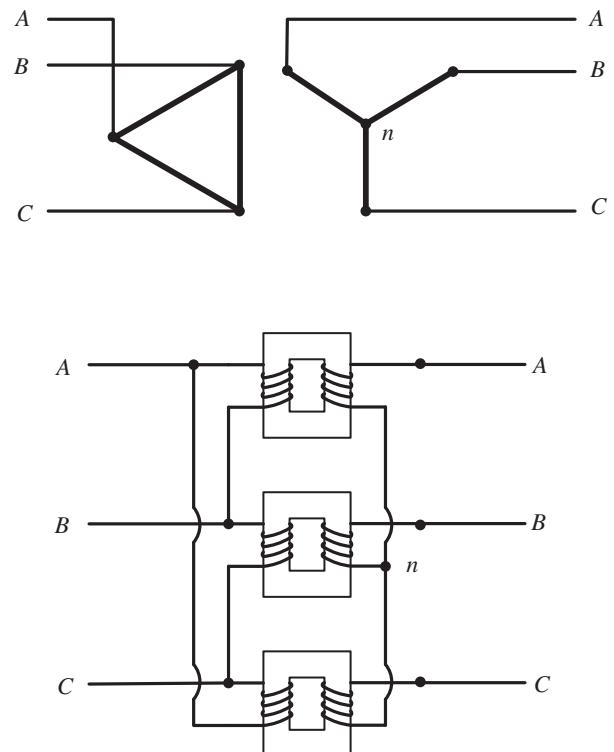


Figure 8.3 Three-phase transformer connections.

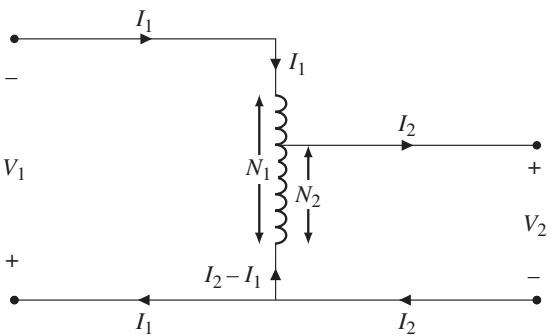
phase-to-ground voltage, $230/\sqrt{3} = 133$ kV. Due to the turns ratio, the voltage across the secondary transformer winding is less by a factor of 10, or 13.3 kV. Because of the delta connection on the secondary side, this is the phase-to-phase voltage, and 13.3 kV is the correct label for the secondary voltage. All in all, the voltage has been stepped down by a factor of $10 \cdot \sqrt{3} = 17.32$.

Beyond the change in the voltage magnitude, the delta-wye conversion in transformers also affects the phase angle. From Figure 4.4, we can appreciate that there is a change in the timing when we compare the phase-to-neutral with the phase-to-phase voltage. The difference, as can be determined through trigonometry or by inspection of the graph, is 30° ($\pi/6$ radians).

This could pose a problem if two parallel paths in a network had a different phase shift. In this case, two voltages would meet again after having gone through their respective transformers, one shifted 30° relative to the other. Such an addition within a network loop creates a circulating current, which is wasteful and possibly dangerous. Therefore, power system designers try to ensure that any pair of parallel paths in a network have equal transformer gains in terms of both voltage and phase. A system in which this is true is termed a *normal* system.

8.4 Autotransformers

The *autotransformer* is a variant that offers simplification and cost reduction at the expense of some operational vulnerabilities. Instead of having separate, electrically isolated sets of windings on the primary and secondary sides, an autotransformer uses a single winding for both voltages.

Figure 8.4 Autotransformer.

It can be conceptualized as an inductive voltage divider, analogous to the capacitive voltage divider illustrated in Figure 16.4 in the context of measurements. Autotransformers typically include a mechanical tap changing feature as introduced in Section 7.4.1, allowing adjustments to the voltage under regular operation.

The basic idea is to tap the single transformer winding at a different place from the primary and secondary side, thus incorporating a different number of turns in the respective circuits. The voltages on either side will be directly proportional to the number of turns connected. For example, as illustrated in Figure 8.4, if the secondary side of the autotransformer includes half the shared winding, the voltage will be stepped down in a 2:1 ratio.

In principle, the step-up and step-down function is completely analogous between a regular and an autotransformer. The inductive coupling with the magnetic field in the transformer core doesn't actually care if the two windings with their respective number of turns are physically separate, or superimposed on each other.

We can deduce from energy conservation that if the voltage is stepped down to half, then the current on the secondary side must accordingly be twice as high, neglecting losses. This is not immediately obvious from the figure, where we must pay attention to the reference direction and handedness of the winding. In the regular simple transformer as shown in Figure 8.1, note that the windings are mirror reflected. Using the right-hand rule, we can see this implies that the instantaneous magnetic flux from the two windings will cancel when their currents are the same in the reference direction (left to right, with one going into the transformer and the other coming out). In the autotransformer in Figure 8.4, where we are tapping the same physical coil with the secondary circuit, we also want to see the flux cancel. This happens naturally when we define the current as positive leaving the tap on the secondary side, where the voltage is positive from top to bottom like the primary side. Applying Kirchhoff's current law (Section 2.3.2) at the tapping point, the current associated with the secondary side circuit is subtracted from that of the primary side when superimposed.⁴ Since flux linkage is the product of current and number of turns (Eq. (8.1)), the flux associated with the primary and secondary windings will be equal and opposite when the secondary current (with half the number of turns) is twice that of the primary.

In an autotransformer, the two sides remain electrically coupled because they share the same neutral and thus a common ground. A regular transformer provides *galvanic isolation* between the two windings, which are coupled only magnetically while their neutral references are independent

⁴ To see that the two designs are consistent with respect to flux, imagine sliding the secondary winding from the regular transformer around the core until it overlaps with the primary. It has opposite handedness, but the current flows from bottom to top while the voltage polarity is upside down. To match it to the autotransformer configuration, we would flip both the handedness and voltage reference direction of the secondary, leaving its flux still opposite that of the primary winding.

and no current is directly shared. The lack of isolation in an autotransformer creates a vulnerability to *ground loops*, or circulating currents due to voltage differences between physically distant points on a circuit that are nominally supposed to be at the same potential. It also means that any disturbance or fault current on either side will directly affect the other. On the other hand, autotransformers require much less copper and steel for the same rated capacity, which makes them less expensive, smaller, lighter, and easier to transport.

8.5 Transformer Modeling

Since a transformer consists of physically separate electric circuits that are coupled by magnetic interaction, it is not obvious how to create a circuit model. How can we account for the way that current and voltage on one side affect current and voltage on the other side, when there are no electrical connections between them? The solution to this modeling problem is to invent an abstraction called the *ideal transformer*, which captures the essential and universal transformer features within a special graphic symbol. We then augment that ideal transformer with electric circuit components that capture the actual behavior of a realistic transformer, in an *equivalent circuit*.

The ideal transformer contains the magic of transferring voltage and current between one winding and the other, stepped up or down by exactly the turns ratio. There are only two pieces of information needed to describe an ideal transformer: the turns ratio, and the polarity. The turns ratio n_1/n_2 is often abbreviated with symbol a .

Polarity simply means that we must pay attention to how the circuit on the secondary side is connected relative to the handedness of the windings, which could flip the voltage and current upside down. The graphic convention is to place a dot at the top or the bottom of each winding. When the dots are aligned, it means that a positive voltage with the same reference direction (e.g., plus on top, minus on the bottom) will be associated with a positive current going in one side and out the other. In other words, if the turns ratio were 1, an ideal transformer with matched polarity markings would completely disappear from the circuit. Aside from turns ratio and polarity, all ideal transformers are identical. They have no losses, and no capacity limitations—just a perfect abstraction.

8.5.1 Nonideal Characteristics

In reality, transformers get warm when they operate; sometimes they buzz audibly. These energy losses are accounted for by resistive elements in the transformer model. There are also inductive elements, which do not consume energy but account for the actual voltage and current being slightly different from the ideal case. Several different physical phenomena come into play, having to do both with the transformer windings and the magnetic core.

The windings of a real transformer will have some nonzero resistance, which is modeled as being in series with the ideal transformer, since the entire load current must travel through it. Accordingly, there will be I^2R losses, sometimes referred to as *copper losses*, because windings are usually made from copper wire.

Real transformer windings have another imperfection called *leakage flux*, which is modeled as a series inductance. This is harder to conceptualize, and a bit confusing because inductance is what a transformer winding is supposed to be all about. The key here is to distinguish between magnetic flux that serves to connect both sides of the transformer—that is, flux that *links* both windings—and flux that doesn't.

To the extent that the flux links both sides, the current flowing in the secondary winding effectively neutralizes the inductance of the primary winding. We observe this effect in that as soon as a load is connected to the secondary side and allows current to flow, a loaded transformer has very little series impedance. By contrast, when the load is disconnected and the secondary winding left open-circuited, the transformer becomes a very high impedance. We know this because even if we leave the primary side connected to a power source, there is not a large current flowing through the primary winding—despite the fact that this winding remains electrically connected. If we didn't know about magnetic effects, the transformer winding would seem to resemble a short circuit. However, its coiled-up shape endows it with an inductance that greatly impedes current flow—until the secondary side is activated to cancel the flux.

This cancellation is not perfect, however. The way this is shown graphically is by drawing flux lines around each winding, and noting that some lines do not intersect the opposite winding. This means that some fraction of the flux does not interact with the other winding, but is left as a residual inductive reactance. We say that the flux “leaks” in the sense that it does not follow a useful path through our device. Again, because the entire transformer current has to travel through the windings and cannot escape its effect, the leakage flux is modeled as an inductive component in series with the ideal transformer.

Like resistance, the inductance associated with leakage flux applies to both primary and secondary sides, though not in equal amounts as the windings will have different lengths. For a complete transformer model, we thus have four series components to specify. In Figure 8.6, these are labeled R_1 , X_1 , R_2 , and X_2 . Since we can add series impedances in any order, it does not matter whether we imagine the resistance and inductance of the primary and secondary windings before or after the ideal transformer. In fact, it will be convenient to combine them as if they were on the same side.

This requires *referring* the series impedance from one side to the other. Referring the impedance addresses the fact that the two sides of the transformer are operating at different voltages. Therefore, the same physical impedance in ohms would have a different effect, depending on whether it was placed on the primary or secondary side. To refer the impedance, we compensate for the change of voltage context by multiplying by the square of the turns ratio, to obtain the equivalent impedance on the other side that would produce the same effect.

When impedances are given in per-unit (see Section 8.7), there is no need to refer them: the same per-unit value would automatically be interpreted as having a different number of ohms, depending on which side of the transformer it was placed (i.e., depending on the base voltage and the base impedance on that side).

Let us now turn to the magnetic core. Its nonideal properties include the fact that magnetization of the material is not complete, which somewhat limits the flux through the core. This is described by a finite core permeability μ_C , or conversely, a nonzero core reluctance \mathcal{R}_C (the magnetic analogues of a finite conductance and nonzero resistance). The physical consequence is that a slightly lesser current will flow in the secondary winding than expected from the turns ratio. This effect is captured in the transformer model as a parallel circuit branch on the primary side, indicating that there is some current in the primary winding that fails to have any impact on the secondary side. Because the effect is small, it is represented by a small parallel (shunt) admittance. This admittance is inductive because the current still has to flow through the coil-shape windings and therefore lags the applied voltage.

Another set of core properties are modeled as a shunt conductance. These represent all the ways in which energy is dissipated in the magnetic core, ultimately turning to heat. Because most

transformer cores are made from iron, the associated losses are called *iron losses*. The physical phenomena responsible include eddy currents, hysteresis, magnetostriction, and mechanical vibration.

Eddy currents arise from the fact that iron conducts electricity, and the core has a solid three-dimensional shape that gives ample opportunity for internal loops of current to flow in various directions, induced by magnetic fields in their neighborhood.⁵ Because eddy currents do not translate into an organized electromotive force on the secondary winding, their net effect is to apply some of the energy from the primary winding to produce waste heat in the core. This loss is modeled as a shunt conductance because it dissipates energy and slightly reduces the available current on the secondary side.

Hysteresis describes the fact that as an applied electric current exerts a magnetizing force on the core material, the microscopic magnetic domains inside (i.e., groups of atoms with a directional spin producing a north and south pole) do not instantly and effortlessly realign. With alternating current at 50 or 60 Hz, the direction of magnetization is being reversed 100 or 120 times each second. However, because the core retains some memory of its previous state of magnetization, it does not respond linearly to the changing external field.

Figure 8.5 illustrates the phenomenon of hysteresis. For a given current applied, which represents a particular magnetic field strength H , the magnetization M of the core depends on its history of exposure, or the path by which it arrived at its present state. If a magnet starts out in a completely neutral state (i.e., the internal magnetic domains point in random directions) then application of an increasing H field will produce a roughly proportional magnetization M . At some point, just about all the domains will be aligned as well as they can be and the magnetization no longer increases; we say the core is *saturated*.

It is not desirable to reach the point of saturation, in part because we would essentially be wasting current without putting it to good magnetic use. Worse, saturation will alter the shape of the output current waveform, manifesting as harmonics that cannot produce useful work and end up producing unwanted heat. That said, if studied under the oscilloscope, many commercial power transformers—especially less expensive ones—can be seen to exhibit saturation. Ideally, however, the alternating current will decrease and begin to reverse before the point of saturation is

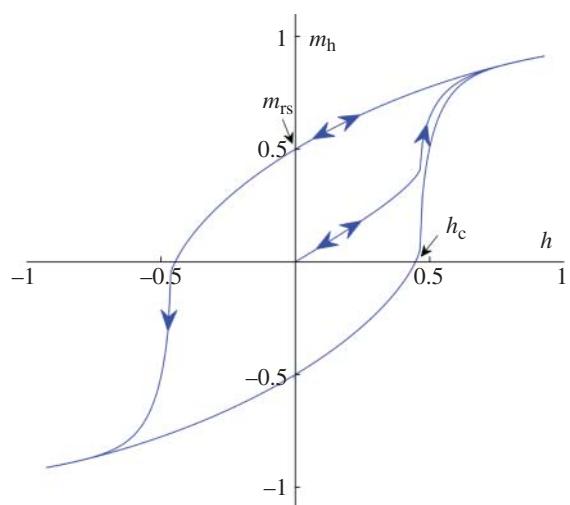


Figure 8.5 Magnetization versus magnetic field strength for a theoretical transformer core, illustrating hysteresis. The intercepts are m_{rs} for saturation remanence and h_c for coercivity. Source: RockMagnetist/Wikimedia Commons/Public Domain.

⁵ This phenomenon is much reduced by grain-oriented electrical steel, which contains silicon for lowering electrical conductivity and has a built-in directionality that favors magnetic fields in the desired direction.

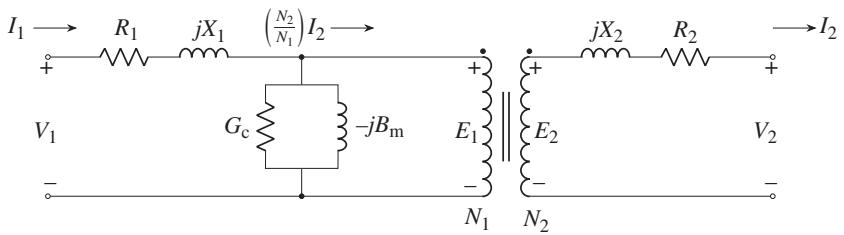


Figure 8.6 Equivalent circuit model for a realistic transformer.

reached. This assumes the core is large enough, so that there is plenty of material to magnetize. At higher frequencies, where current is not sustained for as long in each direction, transformers can be made with smaller cores.⁶ It is generally okay to operate a transformer at a higher frequency than what it was designed for (e.g., deploying a 50-Hz transformer in a 60-Hz grid), but not *vice versa*, due to core saturation.

As the current decreases, since the core was just recently near its maximum magnetization, it will retain more of that magnetization than would be associated with the instantaneous H field had it started from zero. Even as the current crosses zero and the field changes direction, some of the previous magnetization remains (in an amount called *remanence*). The H field will have to get much stronger in the new direction before it finally coerces the magnetization to switch (at the point called *coercivity*). A continually increasing current will now increase magnetization with the new polarity until the a.c. cycle reverses again and repeats the process. Throughout this cycle, the state of the magnetic core describes a *hysteresis loop*. The fact that this is not a reversible back-and-forth path indicates that physical work is being done on the material, where the area enclosed by the loop can be interpreted as a measure of energy dissipated.

All this is more detail than needed for a basic electrical circuit model of a transformer. Here, we can simply capture the energy loss due to hysteresis in terms of a small shunt conductance, which produces the net effect of slightly heating the core. However, it is important to understand that a transformer circuit model assumes a specific operating frequency (such as 50 or 60 Hz) and would have to be revised to characterize the same transformer's behavior when connected to a different source. A related fact is that core saturation is not captured by a standard transformer model, although it can have significant practical impact.

Another mechanism of energy loss is the phenomenon of *magnetostriction*, where a magnetic material deforms ever so slightly as the polarity reverses. This does work on the material at the microscopic level, producing heat. Finally, the alternating magnetic forces on the macroscopic scale will cause the entire transformer to vibrate slightly, often with an audible hum. The small amount of energy consumed by all of these non-idealities collectively is represented by the shunt conductance in the transformer model. When a less accurate model is needed, the shunt elements may be neglected and only the series elements included, as in Figure 8.7.

8.5.2 Referred Impedance

Figure 8.6 illustrates the complete equivalent circuit based on the above discussion. It centers on an ideal transformer, augmented by complex shunt admittance and series impedance. The figure shows a series impedance on both sides, one for each winding, but we will combine these terms by

⁶ A smaller core means not only a cheaper but also a lighter transformer. This is one reason why a standard frequency used in aircraft, where weight matters a lot, is 400 Hz.

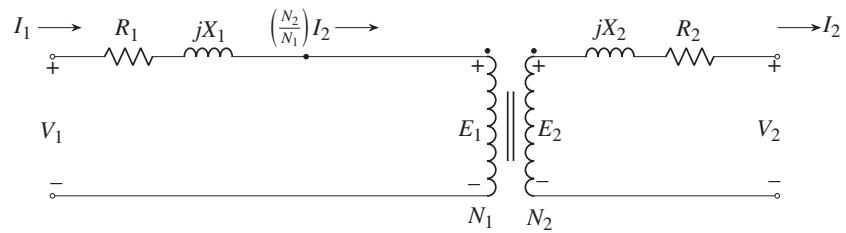


Figure 8.7 Transformer example with series impedance only, neglecting shunt admittance.

referring one to the other side and adding them. This reflects the physical behavior, where series impedances are additive, and simplifies the model.

There is only one shunt admittance, because there is only one core. Note that it does not matter on which side of the ideal transformer these components are placed, because that ideal transformer is not a physical circuit element; it is an abstraction which does nothing but scale the voltage and current by the turns ratio. We only need to ensure that the impedances are properly scaled to the voltage on the side where they are represented, using Eq. (8.4). Placing the shunt admittance and series impedance on the same side helps avoid confusion.

$$Z'_2 = \frac{E_1}{I_1} = \frac{N_1}{N_2} E_2 \div \frac{N_2}{N_1} I_2 = \left(\frac{N_1}{N_2}\right)^2 \frac{E_2}{I_2} = \left(\frac{N_1}{N_2}\right)^2 Z_2 \quad (8.4)$$

Example

Consider a transformer designed for stepping down 480 to 120 V, with turns ratio 4:1 and a power rating of 20 kVA. Suppose it has a series impedance of \$Z_1 = 0.08 + j0.8 = 0.8\angle84.3^\circ\Omega\$ and \$Z_2 = 0.015 + j0.15 = 0.15\angle84.3^\circ\Omega\$ on the primary and secondary side, respectively, as illustrated in Figure 8.8. We neglect shunt admittance.

To refer the impedance from the secondary to the primary side, we multiply it by the square of the turns ratio.⁷ This answers the question, what impedance, if connected to the primary side voltage, would draw the same amount of power? We denote the referred impedance with a prime. Thus, \$Z'_2 = 4^2 \times 0.15\angle84.3^\circ = 2.4\angle84.3^\circ\Omega\$. The total series impedance relative to the primary side is \$Z_1 + Z'_2 = (0.8 + 2.4)\angle84.3^\circ = 3.2\angle84.3^\circ\Omega\$. Note that the impedance of the secondary windings, although much smaller in absolute terms, has a greater impact on performance in this example.

Conversely, the primary side impedance referred to the secondary side is \$Z'_1 = 1/16 \times 0.8\angle84.3^\circ = 0.05\angle84.3^\circ\Omega\$, and the sum of impedances, scaled to the secondary side voltage,

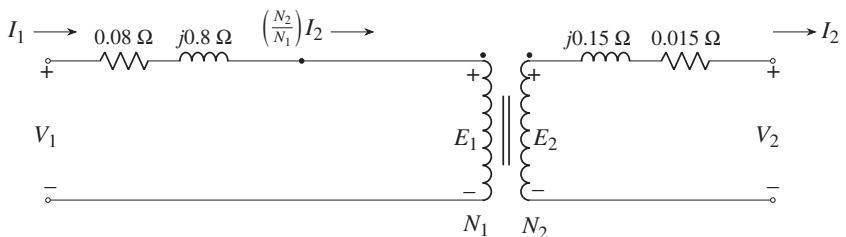


Figure 8.8 Transformer example with series elements only (neglecting shunt admittance).

⁷ The numbers in the example were chosen for convenience with the same ratio of resistance to reactance on both sides, to make polar coordinates easy to use.

is $Z'_1 + Z_2 = 0.2\angle 84.3^\circ \Omega$. Clearly, there is a risk of confusion when expressing transformer impedances in ohms!

This problem motivates the per-unit system, presented in Section 8.7. If the impedances are expressed in per-unit values instead of ohms, we no longer need to worry about which side they are referred to, because per-unit values will be the same on either side.

8.5.3 Open-Circuit and Short-Circuit Tests

It is often adequate for practical purposes to use a simplified transformer model, like in the above example. Some of the most common practical questions about transformers concern the voltage drop as a function of load, and the fault current in case the secondary side is short-circuited. Both of these questions can be answered using only the series impedance, while neglecting the shunt admittance. Furthermore, the series resistance may be small compared to the inductance, in which case a simplified model may include only a series inductance along with the ideal transformer. Note that these levels of simplification are analogous to those used for transmission line models in Section 9.3.

Nevertheless, it is instructive to contemplate how we would come up with a complete equivalent circuit for a real transformer, if needed. There are two specific physical experiments that can be performed to determine the equivalent model parameters: one for the series and one for the shunt component. These experiments are called the short-circuit and the open-circuit tests, respectively.

In the short-circuit test, the secondary side is short-circuited with zero load impedance. If the transformer were ideal, this would result in an arbitrarily high current—as much current as the actual source on the primary side can supply. But here we assume the voltage source is close to ideal over the test range, and the series impedance of the realistic transformer is what limits the current. We ignore the shunt admittance. The series impedance will be approximated by the ratio of voltage to current in the primary winding.⁸

For determining this ratio experimentally, it is tempting to propose we connect the transformer to the rated primary voltage, and then measure the current. On second thought, though, that setup could easily damage our transformer, because the current might be very large and cause the primary windings to overheat. A much better idea is to connect the transformer to a variable voltage source and gradually increase the voltage, while carefully watching the current until it reaches the value of the transformer's full current rating at 100% load. This is not only safer, but more representative of the transformer's typical operating condition (recognizing that its impedance at extremely high current and temperature may change). We then simply read off the voltage required to produce the rated current. Note that the ratio of voltage and current magnitude will give us the impedance magnitude only. If we want to know its composition in terms of resistance and reactance, we must also measure the phase angle between current and voltage.

For the open-circuit test, the secondary side is left open, with no load attached.⁹ This will prevent any current from flowing in the secondary winding. When a voltage source is connected to the

⁸ The procedure is not exact, because we are neglecting the shunt admittance. However, it is probably fair to assume that the shunt admittance representing nonidealities must be orders of magnitude smaller than the admittance of the transformer's windings, whose job it is to conduct current. Therefore, accounting for the shunt admittance should not change the total current much when there is a load, but it would make our experimental setup and calculation much more difficult.

⁹ In practice, it is more convenient to perform the test on the low-voltage side, which is usually labeled as the secondary side, and open-circuit the primary. As far as the transformer model goes, the results should be equivalent. Our description here continues with the test backwards from common practice, for conceptual clarity next to the short-circuit test.

primary side, whatever current flows in the primary winding must be accounted for by the shunt admittance. Because we expect this admittance to be small—that is, the impedance is high—it is now safe to supply the full rated voltage, and measure the corresponding current on the primary winding. In this test, we neglect the series impedance, which we reasonably expect to be orders of magnitude less than the shunt impedance that is effectively throttling the current. Again, if we measure current and voltage magnitudes only and not angles, we will get a magnitude for the shunt admittance but no composition in terms of real and imaginary parts.

Keep in mind that the real transformer has no actual separate circuit branch like the equivalent circuit model in Figure 8.6. What we are teasing out in the open-circuit test is the extent to which the transformer permits a current in one winding that is decoupled from the other winding. It is our faith in the “equivalence” of the circuit model that let us use this shunt admittance to predict the excess current which, under normal operation, acts to magnetize and heat up the transformer.

8.6 Voltage Regulation

Whenever there is a series impedance in the chain of power transmission and distribution, it will cause a voltage drop that varies with current, based on Ohm’s law. The term *voltage regulation* in this context refers not to any active control (as in Section 7.4), but to the unintentional decline of voltage at the receiving end, due to the nonideal character of the devices between generation and load.

Voltage regulation in this sense is defined by the percentage difference between the receiving end voltage magnitude at full load, $|V_{FL}|$, versus no load, $|V_{NL}|$. Values up to about 5% are common. At zero load, receiving end voltage will be the same as sending end. The voltage regulation may be specified for any individual device such as a transformer or a transmission line, or for a combination of series components. In equation form, voltage regulation in percent is given by

$$VR = \frac{|V_{NL}| - |V_{FL}|}{|V_{FL}|} \times 100\% \quad (8.5)$$

With typical numbers on the order of a few percent, it doesn’t make much difference whether one uses the no-load or the full-load voltage in the denominator; by convention, we take V_{FL} as the reference.

It is important to note that Eq. (8.5) uses all scalar quantities and does not involve phasors. However, in order to properly calculate the voltage drop from a given impedance in the transformer or transmission line model, it is necessary to apply Ohm’s law and Kirchhoff’s voltage law (KVL) with phasor quantities. As illustrated in Figure 8.9, the voltage drop itself is the *vector* difference $V_{NL} - V_{FL}$ in the complex plane, which is not the same as the difference in the lengths $|V_{NL}| - |V_{FL}|$.

Example

Consider the transformer from Figure 8.6 and the previous example. How much voltage regulation does it have?

We will determine the voltage drop at full load using only the series impedance (neglecting the effect of shunt admittance). The transformer’s total series impedance calculated above for the primary side is $Z_{tot,1} = 3.2\angle84.3^\circ\Omega$. Although the answer comes out the same, it is typical to use quantities referred to the secondary side when calculating voltage regulation. In our case, we have $Z_{tot,2} = 0.2\angle84.3^\circ\Omega = 0.0199 + j0.199\Omega$.

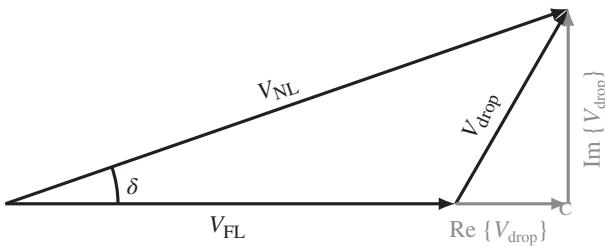


Figure 8.9 Vector addition for voltage drop across a series impedance, in the case where load current is aligned with voltage (p.f. = 1.0).

To determine the voltage drop caused by this series impedance, we must multiply it by the current at full load. On the secondary side, the current at rated power is $I_{FL,2} = 20,000 \text{ VA}/120 \text{ V} = 166.7 \text{ A}$ (and we will drop the subscript 2 from here on).¹⁰ But we also need to know the load power factor.

Suppose that the power factor is unity. We can choose the angle of both the full-load current and full-load voltage to be zero, to write $I_{FL} = 166.7\angle 0^\circ \text{ A}$. In that case, the voltage drop is

$$V_{drop} = I_{FL}Z_{tot} = 166.7\angle 0^\circ \text{ A} \cdot 0.2\angle 84.3^\circ \Omega = 33.34\angle 84.3^\circ \text{ V} = 3.31 + j33.1 \text{ V}$$

The angle of the voltage drop is important because we must subtract it from the no-load voltage in the complex plane in order to get the correct magnitude (even though the voltage regulation formula will just use the comparative magnitudes). KVL is easiest to apply graphically. The vector addition should give $V_{FL} + V_{drop} = V_{NL}$. As for the no-load voltage, we know only the magnitude, $|V_{NL}| = 120 \text{ V}$.

Using trigonometry in Figure 8.9 (which is completely analogous to Figure 9.10) we can determine that V_{NL} must lead V_{FL} by $\delta = \sin^{-1}(\text{Im}\{V_{drop}\}/|V_{NL}|) = 16^\circ$, and that $V_{FL} + \text{Re}\{V_{drop}\} = V_{NL} \cos \delta = 115.3 \text{ V}$. Thus, $|V_{FL}| = 115.3 \text{ V} - 3.31 \text{ V} = 112 \text{ V}$.

Using this value in Eq. (8.5), we find the voltage regulation at unity power factor is

$$\text{VR} = \frac{|V_{NL}| - |V_{FL}|}{|V_{FL}|} \times 100\% = \frac{120 \text{ V} - 112 \text{ V}}{112 \text{ V}} \times 100\% = 7.1\%$$

Let us re-emphasize is that the scalar difference in the numerator is not the same as the actual voltage drop magnitude, which was a whopping 33 V in this example. Owing to the inductive reactance of the transformer and the load's unity power factor, the voltage drop occurs mainly in the imaginary direction, meaning that it represents mainly a shift in time rather than rms magnitude. The voltage regulation is very sensitive to the load power factor because the current angle determines the direction of the voltage drop in the complex plane.

8.6.1 Approximation

The voltage regulation across a device (transformer or line) with given impedance $Z = R + jX$ and a given load displacement power factor $\cos \theta$ can be approximated by Eq. (8.6), shown here for the case of lagging power factors (negative θ):

$$V_R = \frac{IR \cos \theta + IX \sin \theta}{|V_{FL}|} \times 100\% \quad (8.6)$$

¹⁰ We could have used the primary side current, along with the series impedance referred to the primary side. While the nominal voltage drop at full load would be greater by a factor of four on the primary side, it would be the same on a percentage or per-unit basis.

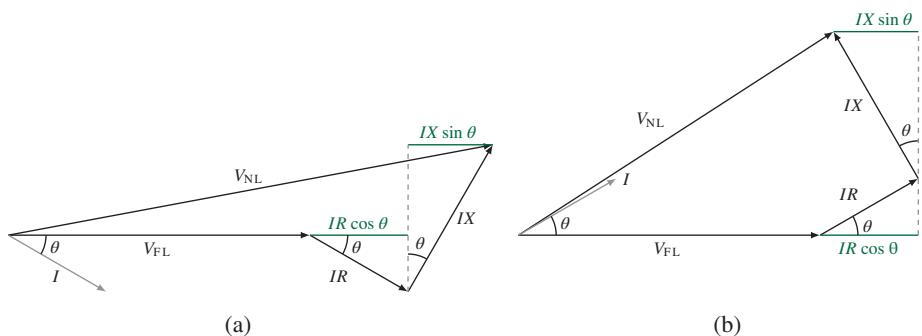


Figure 8.10 Projecting voltage drop onto the real axis to approximate $|V_{NL}| - |V_{FL}|$ as in Eq. (8.6), for lagging and leading current. Voltage drop is exaggerated for clarity.

For the case of a leading power factor (positive θ), the plus sign in the numerator becomes a minus sign.

This approximation works nicely in situations where the angle δ is small, meaning that much of the voltage drop is in the real direction (drawn horizontally, in line with the full-load or receiving-end voltage). It is illustrated in Figure 8.10, which should be compared to Figure 9.10. The expression in the numerator of Eq. (8.6) includes the real (horizontal) projections of the product of current with R and X , respectively. In the case where the no-load or sending-end voltage is pointing roughly in the same direction as the full-load or receiving-end voltage, this real projection approximates the difference in magnitudes of the two voltages.

Note that this approximation does not work well for the previous transformer example, where the voltage drop is mostly in the imaginary direction, so δ is large.

8.7 Per-unit System

The per-unit (p.u.) system offers a way to express electrical quantities relative to their environment that is meaningful and comparable across different situations. A per-unit value of voltage, current, power or impedance is some fraction of the *base* quantity that typically describes a nominal or full rated operating condition. For example, we would expect a load of 0.1 p.u. to represent 10% of the full rated power for the circuit.

This system helps deal with transformers, where the voltages and currents on either side may be of an entirely different order of magnitude. A load with a certain impedance in ohms that draws some number of amperes on the primary side would draw a very different number of amperes on the secondary side. This is why it becomes necessary to *refer* impedances, to scale them appropriately for the voltage level at which their effect is being considered. However, the per-unit impedance and current are essentially the same on either side of the transformer, as are per-unit voltage and power. The per-unit system makes ideal transformers disappear completely, while highlighting the quantities associated with nonidealities (such as voltage regulation) that are of practical interest.

There are four base quantities in the per-unit system: two of them, power and voltage, are deliberately chosen by the analyst; the other two, current and impedance, follow directly from that choice. While the choice of base power and voltage is, mathematically speaking, arbitrary, the *convention* is to use equipment power ratings and nominal voltages. This convention ensures that the per-unit values will be intuitively and practically meaningful.

The base power S_{BASE} is chosen to be consistent with the power rating in kVA or MVA of the equipment being described. If there are multiple pieces of equipment with different ratings on the same circuit, one of them is chosen arbitrarily to serve as the reference. Note that we would expect the ratings to be fairly similar, since it does not make physical sense to connect equipment of vastly different capacities together. Regardless of how it was chosen, the base power S_{BASE} will remain the same throughout the analysis for all connected components on the system. Per-unit real, reactive and apparent power will all be specified on the same kVA base.¹¹

The base voltage V_{BASE} is chosen for each *zone* of a system, where different voltage zones are defined by the transformers that separate them. The base voltage on the primary and secondary side of a transformer will be different, generally in accordance with the turns ratio—this is the key point of the per-unit system. Occasionally, there could be a slight discrepancy in the nominal voltage rating of two transformers connecting to the same zone, in which case one of them is chosen arbitrarily as the reference.

With base power and voltage defined, base current and impedance are given by the following relationships that are consistent with dimensional analysis:

$$I_{\text{BASE}} = \frac{S_{\text{BASE}}}{V_{\text{BASE}}} \quad \text{and} \quad Z_{\text{BASE}} = \frac{V_{\text{BASE}}^2}{S_{\text{BASE}}} \quad (8.7)$$

Base current and base impedance will be different in each zone, since base voltage is different, while base power is the same everywhere.

Example

A transformer with turns ratio $n = 15:1$ is rated for 7.20 kV on the primary side and 480 V on the secondary, with a power rating of 100 kVA.

When a load with impedance $Z_2 = 10.0 \Omega$ is connected on the secondary side, it draws $I_2 = 47.6 \text{ A}$ at an actual operating voltage of $V_2 = 476 \text{ V}$, with a power factor of 0.95 lagging.

We choose the per-unit base quantities as follows: $S_{\text{BASE}} = 100 \text{ kVA}$, $V_{\text{BASE},1} = 7.20 \text{ kV}$, $V_{\text{BASE},2} = 480 \text{ V}$, which gives $I_{\text{BASE},1} = 13.89 \text{ A}$, $I_{\text{BASE},2} = 208.3 \text{ A}$, $Z_{\text{BASE},1} = 518.4 \Omega$, $Z_{\text{BASE},2} = 2.304 \Omega$.

The load is then described in per-unit as having impedance

$$Z_{2,\text{p.u.}} = \frac{Z_2}{Z_{\text{BASE},2}} = \frac{10 \Omega}{2.304 \Omega} = 4.34 \text{ p.u.}$$

and drawing current

$$I_{2,\text{p.u.}} = \frac{I_2}{I_{\text{BASE},2}} = \frac{47.6 \text{ A}}{208.3 \text{ A}} = 0.229 \text{ p.u.}$$

at voltage

$$V_{2,\text{p.u.}} = \frac{V_2}{V_{\text{BASE},2}} = \frac{476 \text{ V}}{480 \text{ V}} = 0.992 \text{ p.u.}$$

Note that the relationship $V_{\text{p.u.}} = I_{\text{p.u.}} \cdot Z_{\text{p.u.}}$ also holds. As another side note, appreciate that if referred to the primary side, the load impedance would appear as

$$Z'_2 = n^2 Z_2 = 15^2 10 \Omega = 2250 \Omega$$

but in per-unit it remains the same:

$$Z_{\text{p.u.}} = \frac{Z'_2}{Z_{\text{BASE},1}} = \frac{2250 \Omega}{518.4 \Omega} = 4.34 \text{ p.u.}$$

¹¹ Recall that watts, VARs and volt-amperes are different labels that all have the same physical dimensions, so the ratio of any one to the other is still dimensionless.

The apparent, real, and reactive power drawn by the load is

$$S_2 = I_2 V_2 = 22.7 \text{ kVA}$$

$$P_2 = 0.95 S_2 = 21.5 \text{ kW}$$

$$Q_2 = \sin(\cos^{-1} 0.95) S_2 = 7.07 \text{ kVAR}$$

or, expressed in per-unit:

$$S_{2,\text{p.u.}} = \frac{S_2}{S_{\text{BASE}}} = \frac{22.7 \text{ kVA}}{100 \text{ kVA}} = 0.227 \text{ p.u.}$$

$$P_{2,\text{p.u.}} = \frac{P_2}{S_{\text{BASE}}} = \frac{21.5 \text{ kW}}{100 \text{ kVA}} = 0.215 \text{ p.u.}$$

$$Q_{2,\text{p.u.}} = \frac{Q_2}{S_{\text{BASE}}} = \frac{7.07 \text{ kVAR}}{100 \text{ kVA}} = 0.0707 \text{ p.u.}$$

If the transformer were ideal, the load power would appear as identical on the primary side, whether physically or in per-unit, and we could omit the subscript 2 on the per-unit quantities. But let's suppose that this is a realistic, somewhat lossy transformer. That information would be implied by giving the primary side voltage as 7.20 kV = 1.000 p.u., while the load is causing a small voltage drop of 0.008 p.u. (just less than 1%) across the transformer.¹² To fully characterize the losses in the transformer, we would also have to know the actual primary side current. Absent any information about shunt admittance in the transformer model (i.e., neglecting core losses), we would assume the primary side current is the same as the secondary side in per-unit.

Example

The impedance of a transformer is specified in per-unit on the nameplate, based on its own power and voltage ratings. When incorporating that transformer impedance in a calculation with other quantities, it is necessary to use consistent base values for each zone.

Figure 8.11 illustrates an example with two transformers T1 and T2 in series that have different power and voltage ratings, although each has a nameplate impedance of 0.05 p.u. Suppose we wish to find the load current, where the source voltage is given as $V_s = 480 \text{ V}$, and the load and line impedance are both given in ohms. In principle, all we need to do is divide the source voltage by

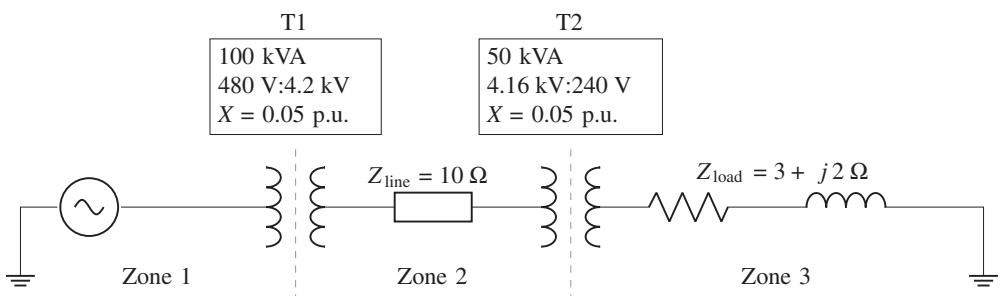


Figure 8.11 Adding series impedances in per-unit across multiple zones with different base voltages.

¹² If we assume that the voltage drop is linear with load, this implies a voltage drop at full rated load of $0.008/0.227 = 0.035 \text{ p.u.}$, or a voltage regulation in the neighborhood of 3.5% (a realistic value).

the combined series impedance. The challenge is to convert all four impedances (T1, line, T2, load) into per-unit values with consistent bases for each zone, so that we may simply add them.

The base power must be the same throughout the system. Let's choose the T1 rating, 100 kVA. This already tells us that the T2 per-unit impedance will have to be adjusted, since it was based on its own rating of 50 kVA.

For the base voltage, there are three different zones. On the left side of T1, which we will call Zone 1, we simply choose its rated voltage $V_{\text{base},1} = 480 \text{ V}$. Conveniently, this lets us set the voltage source to $V_{\text{s.p.u.}} = 1\angle 0^\circ \text{ p.u.}$, and we may keep the T1 impedance of $Z_{\text{T1,p.u.}} = 0.05 \text{ p.u.}$

Zone 2 is the middle section that includes the line between T1 and T2. We would expect this section to be operated at a voltage level somewhere in the neighborhood of 4.1 and 4.2 kV, and we could choose either of these as the base voltage for that zone. Let's make the T1 rating the base, $V_{\text{base},2} = 4200 \text{ V}$. For Zone 3 which includes the load, we choose $V_{\text{base},3} = 240 \text{ V}$.

Our task is now to convert the line and T2 impedances into correct per-unit values for Zone 2, and the load impedance into per-unit for Zone 3.

The base impedance for Zone 2 is

$$Z_{\text{base},2} = \frac{V_{\text{base},2}^2}{S_{\text{base},2}} = \frac{(4200 \text{ V})^2}{100,000 \text{ VA}} = 176.4 \Omega$$

This gives the line impedance in per-unit as

$$Z_{\text{line,p.u.}} = \frac{Z_{\text{line}}}{Z_{\text{base},2}} = \frac{j10 \Omega}{176.4 \Omega} = j0.0567 \text{ p.u.}$$

To find the correct per-unit impedance for T2, it is perhaps least confusing to first find an impedance in ohms (which may be referred to either side; here we choose the high-voltage side), and then express it in terms of the correct base. We have

$$Z_{\text{base,T2}} = \frac{V_{\text{rated}}^2}{S_{\text{rated}}} = \frac{(4100 \text{ V})^2}{50,000 \text{ VA}} = 336.2 \Omega$$

and accordingly

$$Z_{\text{T2}} = j0.05 \text{ p.u.} \cdot 336.2 \Omega = j16.8 \Omega$$

which describes the transformer's primary and secondary side winding impedances jointly referred to the 4.1-kV side. In terms of our chosen base for Zone 2, we get

$$Z_{\text{T2,p.u.}} = \frac{Z_{\text{T2}}}{Z_{\text{base},2}} = \frac{j16.8 \Omega}{176.4 \Omega} = j0.095 \text{ p.u.}$$

The base impedance for Zone 3 is

$$Z_{\text{base},3} = \frac{V_{\text{base},3}^2}{S_{\text{base},2}} = \frac{(240 \text{ V})^2}{100,000 \text{ VA}} = 0.576 \Omega$$

Accordingly, the load impedance in per-unit is

$$Z_{\text{load,p.u.}} = \frac{Z_{\text{load}}}{Z_{\text{base},3}} = \frac{3 + j2 \Omega}{0.576 \Omega} = 5.21 + j3.47 \text{ p.u.}$$

Finally, we may combine series impedances

$$\begin{aligned} Z_{\text{total,p.u.}} &= Z_{\text{T1,p.u.}} + Z_{\text{line,p.u.}} + Z_{\text{T2,p.u.}} + Z_{\text{load,p.u.}} \\ &= j0.05 + j16.8 + j0.095 + 5.21 + j3.47 \\ &= 5.21 + j3.67 = 6.37\angle 35.17^\circ \text{ p.u.} \end{aligned}$$

and the current throughout the entire system is

$$I_{\text{p.u.}} = \frac{V_{\text{s.p.u.}}}{Z_{\text{total,p.u.}}} = \frac{1\angle 0^\circ}{6.37\angle 35.17^\circ} = 0.157\angle -35.17^\circ \text{ p.u.}$$

The current in amps will vary by zone depending on the transformer turns ratios; it can be found simply by multiplying the per-unit current by the appropriate base current for each zone. Crucially, though, the current in per-unit must be the same everywhere along the series path.

Problems and Questions

- 8.1** A step-down transformer is rated 12.0 kV:240 V. Suppose it has 50 turns on the secondary, low-voltage side.
- How many turns are there on the primary side?
 - The same turns ratio could be achieved with many different combinations of n_1 and n_2 . What engineering design factors do you think favor a larger number of turns on both primary and secondary side, versus a smaller number?
- 8.2** A step-down transformer is connected between one phase and neutral on a distribution circuit with a nominal (line-to-line) voltage of 21 kV. The turns ratio of the transformer is 50:1. Neglect any nonideal properties and state your answers to two significant figures.
- What is the rms voltage on the secondary (customer, low-voltage) side of the transformer?
 - Suppose the customer's load is 14.4 kW. What is the current on the primary (high-voltage) side of the transformer?
 - Which conductors need to be thicker: those on the primary or secondary side? Explain.
- 8.3** A single-phase transformer rated 7.4 kVA, 1.2 kV/120 V has a primary winding (on the high-voltage side) of 800 turns.
- How many turns are in the secondary winding?
 - What are the currents in the primary and secondary windings when the transformer is operating at full rated power and rated voltage? Ignore losses (i.e., treat the transformer as ideal).
 - For the following steps, suppose the transformer is delivering 6 kVA at its rated voltages and $p.f. = 0.8$ lagging. What is the impedance Z_2 connected across the 120-V terminals?
 - What is the impedance Z'_2 referred to the primary side? Write a sentence explaining why you would expect Z'_2 to be greater or less than Z_2 .
 - Using Z'_2 , find the primary current and power supplied. Does it agree with your expectations?
- 8.4** A single-phase transformer rated 1.2 kV/120 V, 7.4 kVA has winding resistance $r_1 = 0.8 \Omega$ and $r_2 = 0.01 \Omega$ and leakage reactance $x_1 = 1.2 \Omega$ and $x_2 = 0.01 \Omega$ on the primary and secondary side, respectively.
- Why do you think these quantities are larger on the primary side?
 - Find the transformer series impedance—that is, the combination of winding resistance and leakage reactance from both sides—in ohms, referred to the primary side.
 - Find the transformer impedance in ohms, referred to the secondary side.

- (d) Find the transformer impedance in per-unit.
- (e) If the turns ratio is $a = 10$, what is the actual primary voltage when the transformer delivers 7.4 kVA at 120 V and unity power factor to the secondary side? (Hint: It will be slightly greater than 1.2 kV because the transformer is not ideal.)
- (f) Suppose the primary voltage stays constant at the value you found in (e). But now the load is disconnected, so the secondary current drops to zero. What is the secondary voltage? (Hint: It was 120 V at full load, so at no load it will increase a bit.)
- (g) What is the voltage regulation in percent?
- 8.5** Suppose that the transformer from the previous problem is operating at full rated load when a fault occurs on the secondary side, which reduces the load impedance and the voltage to zero. Assume the only source impedance present is due to the transformer itself, and no protective device trips. What would be the secondary current in this hypothetical situation?
- 8.6** A transformer is connected to an a.c. voltage source on the primary side, while the secondary side has no load connected to it: that is, the secondary winding is an open circuit.
- Suppose this were an ideal transformer. What is the current in the primary winding? Explain in your own words.
 - Now assume this is a practical (real-life), not an ideal transformer. What determines the current in the primary winding when the secondary winding is an open circuit? How is this represented in the transformer model?
- 8.7** What would happen if you accidentally connected a practical transformer to a d.c. voltage source? (*DO NOT* try this at home.)
- 8.8** Imagine you are teaching power engineering and students asks, “What would happen if you connect an ideal transformer to a d.c. source?” How do you respond?
- 8.9** A machine is rated 500 MVA, 20 kV. The windings are Y-connected with a load impedance of 1.1 p.u.
- What are the base quantities for P , Q , S , R , X , Z , and I ?
 - Find the load impedance in ohms.
 - Suppose the same machine is being used in a circuit for which $S_{\text{BASE}} = 200 \text{ MVA}$ and $V_{\text{BASE}} = 22 \text{ kV}$. What is the per-unit value of the load impedance in the new base?
- 8.10** Rework the per-unit example with three zones using $V_{\text{base},2} = 4.1 \text{ kV}$ for Zone 2.
- 8.11** For the per-unit example with three zones, find (a) the I^2X losses in each transformer and on the line in VAR, (b) the total complex power delivered by the source, and (c) the voltage seen by the load.

9

Analyzing Transmission Lines

Transmission lines are surprisingly complicated to analyze, considering that we are basically talking about some straight pieces of metal. Owing to their scale—both in length and cross section—transmission conductors don't simply behave like wires in circuits of more familiar size. Because electric and magnetic fields surrounding a large conductor have a significant effect on its reactance, the physical shape and geometry are important to consider, along with the cross-sectional area and conductivity that determine resistance. Representing all the physical behaviors of a long a.c. transmission line with high fidelity means accounting for inductance and capacitance. We also need to consider that these properties are physically spread out or *distributed* along the length of the line, rather than lumped all in one place. This gives rise to nonobvious behaviors, where voltages and currents vary as a function of distance in a wave-like manner, requiring mathematical descriptions that are not immediately intuitive.

A key theme in transmission line modeling is deciding which of these features are negligible for a particular calculation, and choosing from one of several standard options to simplify the analysis where appropriate. We begin with discussing the basic physical behaviors of inductance and capacitance, before putting them together into complete transmission line models. Note that resistance does not introduce any novel considerations due to the large scale, so the discussion in Chapter 2 is sufficient foundation. A summary observation is that for large power lines, reactive properties tend to be more significant than resistance.

This chapter takes a more mathematical approach than the previous qualitative discussion and can be skipped if desired.

9.1 Transmission Line Inductance

In deriving the inductance of a straight wire, there are two conceptual steps: first, to consider the magnetic field that results from current flowing in the wire, and second, the effect of that magnetic field in terms of *linking* that wire, or creating an inductive effect back onto it. The first step is given by *Ampère law*, which relates the magnetic field surrounding the wire to the current enclosed by it, as illustrated in Figure 9.1.

$$\oint \mathbf{H} \cdot d\mathbf{l} = 2\pi r H = I_{\text{encl}} = I \quad (9.1)$$

The vector variable \mathbf{l} (for line) in Eq. (9.1) describes a circular path around the wire, at a distance r . The equation applies more generally to paths of any shape, in which case the field H might vary along the way. It also applies to situations where the current is spread out in space, at

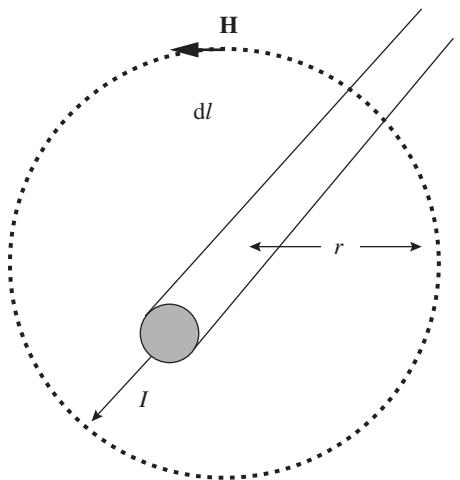


Figure 9.1 The magnetic field around a single current-carrying wire.

some density that may vary. Here, we assume all the current is flowing inside the metal wire. As long as our integration path stays outside the wire, we can drop the subscript denoting *enclosed* for simplicity and just refer to the total current I carried by the wire. We can say by symmetry that H will stay the same along the circle at a constant distance r .

To find inductance, we will write an expression for the magnetic flux linkage λ and then simply divide by current, since

$$\lambda = LI$$

The flux linkage λ is essentially a measure of magnetic flux Φ that is in a position to interact with a particular object—in this case, surrounding the wire.

Recall from Section 2.6 that the magnetic field B represents the flux density (flux per unit area), and that this includes the magnetization of the material exposed to the H -field depending on its magnetic permeability μ . Since our wire is suspended in air, we will use the permeability of space, $\mu_0 = 4\pi \times 10^{-7}$ H/m. Substituting and solving Eq. (9.1), we write:

$$B(r) = \mu_0 H(r) = \frac{\mu_0 I}{2\pi r} \quad (9.2)$$

Equation (9.2) describes the magnetic flux density B at distance r from the wire. To expand this into the total flux *linkage*, we need to multiply by the cross-sectional area. Because B varies continuously as a function of distance, we must take an integral:

$$\Phi = \int_A \mathbf{B} \cdot d\mathbf{A}$$

where the vector dot product indicates that we only count the component of the B field that intersects (i.e., is *normal* to) the area A . Figure 9.2 shows a visualization of flux lines tangent to the wire, with a hypothetical small rectangular area dA .

One side of this rectangle is parallel to the wire. Because the field or flux will not change along the length, we will completely suppress this dimension, by expressing all flux linkage and inductance quantities *per unit length* of the wire.¹ This not only simplifies the expression, but will also yield a nice practical metric at the end, in henrys per meter or per mile.

¹ This assumes that each bit of wire length should contribute equally to its inductive property. If the wire gauge were to vary along the way—say, because of a splice—our calculation would come out a tiny bit wrong. It also means we ignore anything that might happen at the ends of the wire.

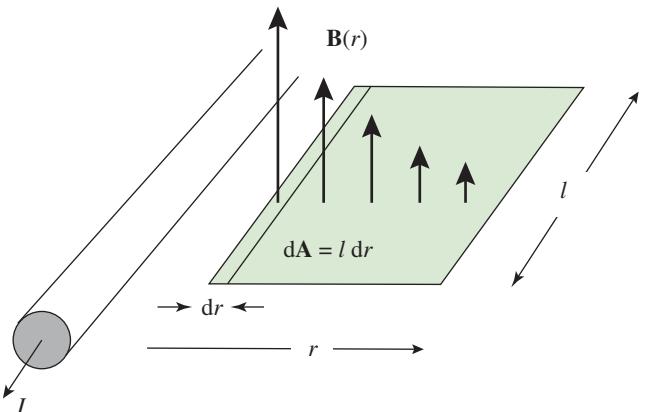


Figure 9.2 Integrating over the magnetic flux density around a single current-carrying wire.

The other side of the small rectangle, labeled dr , extends in the radial direction from the center of the wire outward. As Eq. (9.2) tells us, the flux will vary inversely with distance r . Thus, to take into account the combined effect of magnetic flux at various distances from the wire, we must explicitly add up the different contributions—that is, set up an integral—ranging over all relevant distances r .

Let us consider the flux contribution λ_{12} between some radial distance D_1 and D_2 . Recalling some introductory calculus,² we obtain an expression with a natural logarithm:

$$\lambda_{12} = \int_{D_1}^{D_2} B(r) dr = \int_{D_1}^{D_2} \frac{\mu_0 I}{2\pi r} dr = \frac{\mu I}{2\pi} \ln \frac{D_2}{D_1} \quad (9.3)$$

Now, what to use for the limits of integration D_1 and D_2 ? It is tempting to start from the wire's surface, but in fact we also need to take into account flux *inside* the metal conductor. It turns out that this contribution from the inside can be accounted for by simply using an adjusted version of the wire radius for D_1 , but the rationale is not at all obvious. We will first consider the inside, then the outside region of the wire before we combine the two contributions to flux linkage.

9.1.1 Internal Flux Linkage

The flux linkage inside the conductor is tricky to think about, since the enclosed current will vary as a function of distance from the center. We assume here that this current is spread out evenly over the cross section of the wire. Specifically, the enclosed current $I(r)$ will now be a fraction of total current I in the wire, corresponding to the fraction of cross-sectional area enclosed at radius r , as illustrated in Figure 9.3.

This assumption of uniformly distributed current over the wire's cross section is actually wrong, but in a way that will not substantively affect the conclusions of this section. It is worth a digression because it is conceptually interesting.

Alternating current involves a *skin effect*, which tends to drive more current toward the conductor's surface. This tells us that the current density will *not* actually be uniform. In fact, the magnetic field internal to a conductor, as considered in the present section, explains why the skin effect exists in the first place: because at smaller radii, there is greater flux linkage with the surrounding current, just by virtue of being exposed to more conductor all around. The back-emf associated

² $\int_a^b \frac{1}{x} dx = \ln b - \ln a = \ln \frac{b}{a}$.

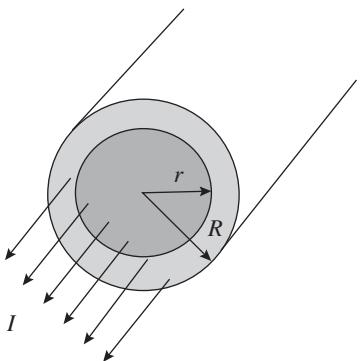


Figure 9.3 Accounting for the fraction of enclosed current at $r < R$ inside the conductor, assuming uniform current density.

with this flux linkage—essentially a self-inductance—thus presents a greater impedance to currents flowing near the center of the conductor than near the surface.

Consequently, the current density decreases exponentially from the conductor surface toward its center. The depth (measured from the surface) where the current density has fallen by a factor of $1/e$ (about 0.37) is called the *skin depth* δ .³ The skin depth decreases for increasing a.c. frequencies, making the skin effect more pronounced.

The approximate formula for skin depth at reasonably low frequencies is

$$\delta = \sqrt{\frac{2\rho}{\omega\mu}} \quad (9.4)$$

Plugging in values for copper, resistivity $\rho \approx 0.017 \times 10^{-8} \Omega\text{-m}$ and magnetic permeability $\mu \approx 1.26 \times 10^{-6} \text{ H/m}$, we get $\delta \approx 0.0085 \text{ m}$, or about one-third of an inch at 60 Hz. For a large conductor, this can have a noteworthy impact on its effective a.c. resistance, since the cross-sectional area that mostly gets utilized by moving charges is less than what we'd expect from its diameter. For this reason, tables of transmission conductor properties often list a d.c. resistance and a (slightly greater) 50 or 60-Hz a.c. resistance.

Ironically, though, for purposes of estimating the contribution of the conductor's internal magnetic fields to its overall inductance, we can get away with ignoring the skin effect.⁴ Besides, the following geometric derivation has such an elegant and appealing result that it's hard to resist.

Proceeding under the assumption of uniform current distribution, if the wire radius is R , the enclosed current at radius r is

$$I_{\text{encl}}(r) = \frac{\pi r^2}{\pi R^2} I$$

This fraction comes into play twice: once because the magnetic field or flux density at distance r is caused only by the enclosed current, and once again because that flux is *linking* only the current inside it. In other words, the flux at distance r has no inductive association with any current flowing farther outside (at radii greater than r) because from the vantage point of that larger radius, the flux inside simply describes a closed circle that isn't being intersected in any way. Thus, we adjust $B(r)$ to include only the enclosed current $I(r)$. We also drop the subscript on the magnetic permeability,

³ No relation to other uses of lowercase delta in this book.

⁴ The ambitious reader may contemplate the direction of the estimation error due to assuming uniform current density.

since μ would now depend on the conductor material.⁵

$$B(r) = \frac{\mu I_{\text{encl}}(r)}{2\pi r} = \frac{\mu}{2\pi r} \frac{\pi r^2}{\pi R^2} I \quad (r < R) \quad (9.5)$$

We now modify Eq. (9.3) using the revised version of $B(r)$, including the factor for the fraction of enclosed area again to describe the flux linkage λ_{int} internal to the wire:

$$\lambda_{\text{int}} = \int_0^R \frac{\pi r^2}{\pi R^2} B(r) dr = \int_0^R \frac{\pi r^2}{\pi R^2} \frac{\mu}{2\pi r} \frac{\pi r^2}{\pi R^2} I dr = \int_0^R \frac{\mu I}{2\pi} \frac{r^3}{R^4} dr \quad (9.6)$$

Remarkably, when evaluating this integral,⁶ the wire radius R cancels, and the internal flux linkage turns out to be just a constant times the total current:

$$\lambda_{\text{int}} = \frac{\mu I}{8\pi} \quad (9.7)$$

It will be easy later to add this internal flux linkage to the external contribution.

9.1.2 External Flux Linkage

For the region outside the wire, we revisit Eq. (9.3) and start with the wire radius R as the lower limit of integration. But what to use for the upper limit, D_2 —do we go all the way out to infinity? In principle, if our current-carrying wire were alone in the universe, then yes, we'd have to integrate B out to infinity. Clearly, this causes problems: the logarithm in Eq. (9.3) will not grow as fast as D_2 , but it will still keep increasing. Even though the magnetic field drops off with distance as $1/r$, the sum of contributions from ever larger distances won't converge to a finite value. What this is telling us is not only that a lone wire has a surprisingly large inductance, but that there is some conceptual flaw in the way we set up the problem. The flaw is that our current-carrying wire cannot be alone in the universe. Sooner or later, all that moving charge needs to return somewhere!

In physical reality, a sustained current—and remember that all this analysis is intended to describe a steady state—requires a complete circuit, with a return flow. Sometimes, if no metal conductor is provided for this purpose, the return current can go through the earth. In any case, that return flow will occur at some finite distance, and it will have an important impact on the magnetic field. In fact, it only makes practical sense to describe the inductance of a complete power line that includes the entire ensemble of conductors, whether it is just a pair of wires for a single-phase line, or a trio of three-phase conductors, or some other combination.

What we find when we evaluate the magnetic field or flux at some point P , a distance r away from the original wire under consideration, is that there will necessarily be another contribution to the magnetic field or flux at that point, due to one or more other wires in the vicinity that carry the return current. Now, since the return current goes in the opposite direction, the magnetic field or flux lines will tend to cancel each other. The formal way to see this is to go back to Ampère's Law in Eq. (9.1), and note that the current enclosed by a path integral around multiple conductors with opposite currents will be zero.⁷

Let's consider a single-phase circuit, where the two conductors carrying exactly opposite currents are strung at a constant spacing D from each other, as shown in Figure 9.4. If we go a very large

⁵ Not all electrical conductors are magnetic; for aluminum, the relative permeability is close to 1 (i.e., the same as air or empty space). Steel can have a higher value of μ , but the effect in this context is small and rarely considered.

⁶ $\int_0^R r^3 dr = \frac{1}{4}R^4$.

⁷ This is why you cannot measure current by clamping a current transformer around a twisted pair of wires that include both legs of the circuit: it can't pick up the magnetic field from one individual wire.

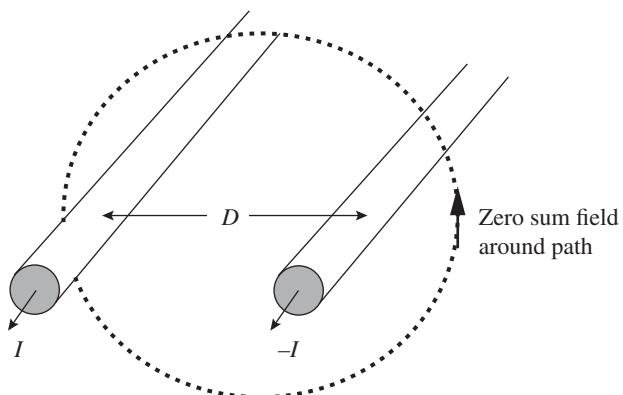


Figure 9.4 When the return conductor is inside the path integral, the enclosed current and therefore the magnetic flux is zero. The same applies to a three-phase circuit.

distance $r \gg D$ away from the circuit, the field cancellation will be nearly perfect since the two wires will appear about equally far away and so will produce equally strong fields, in opposite directions. For points directly in between the two wires, the fields will actually tend to add by right-hand rule, but the vector sum at any given point in space is not obvious. Fortunately, the value of the magnetic field at any one point in space is not what counts because the fields from all points around the conductor contribute in various amounts. The effect is summarized by the flux *linkage* that causes induction—that is, the field or flux lines that wrap completely around the conductor in a closed path.⁸ This means the path integral from Ampère's Law is the correct way to account for the flux. Conveniently, if we draw a circular path of radius r around either one of the wires, the result is simple: if $r < D$, the path integral of the field must come out to match the current on the one enclosed wire; the fact that the field of the other wire added a bit on one side and subtracted a bit on the other is completely accounted for. On the other hand, if $r > D$, the path integral must yield zero, despite encountering nonzero fields along the way. Therefore, the distance D to the return conductor is exactly the upper limit of integration D_2 in Eq. (9.3) we are looking for: beyond D_2 , there will be no further contributions.

For a single-phase circuit, we can write the external contribution to the flux linkage as

$$\lambda_{\text{ext}} = \int_R^D B(r) dr = \int_R^D \frac{\mu_0 I}{2\pi r} dr = \frac{\mu I}{2\pi} \ln \frac{D}{R} \quad (9.8)$$

9.1.3 Per-Phase Inductance

With this, we are ready to combine the internal and external contributions to flux linkage. We will neglect any discrepancy between the magnetic permeability of the metal and air, and assume $\mu = \mu_0$. From Eqs. (9.7) and (9.8), we have:

$$\lambda_{\text{total}} = \lambda_{\text{int}} + \lambda_{\text{ext}} = \frac{\mu I}{8\pi} + \frac{\mu_0 I}{2\pi} \ln \frac{D}{R}$$

⁸ This requirement for continuity explains the use of the word “flux,” as if there were a material substance (subject to conservation, like mass conservation) flowing through space.

With some clever rearrangement, the added constant gets tucked away into the argument of the logarithm, and we substitute the numerical value for the term with μ_0 :

$$\begin{aligned}\lambda_{\text{total}} &= \frac{\mu_0 I}{2\pi} \left(\frac{1}{4} + \ln \frac{D}{R} \right) = \frac{\mu_0 I}{2\pi} \left(\ln e^{\frac{1}{4}} + \ln \frac{D}{R} \right) = \frac{\mu_0 I}{2\pi} \ln \frac{D}{e^{-\frac{1}{4}} R} \\ &= 2 \times 10^{-7} I \ln \frac{D}{0.7788 R} = 2 \times 10^{-7} I \ln \frac{D}{R'}\end{aligned}\quad (9.9)$$

Because the factor 0.7788 occurs so often in these types of calculations, it is conventional to define an adjusted, effective conductor radius R' .⁹

$$R' \equiv e^{-\frac{1}{4}} R \approx 0.7788 R$$

Now, since the inductance is simply the ratio of flux linkage to current, we only have to cancel the current I to get an expression for the inductance per-phase L_p of a single-phase circuit, per unit length:

$$L_p = \frac{\lambda_p}{I_p} = 2 \times 10^{-7} \ln \frac{D}{R'} \text{ (H/m)} \quad (9.10)$$

Because the value of μ_0 is expressed in SI units, the result is in henrys per meter. The entire single-phase circuit would have twice this inductance, to account for the return conductor as well.

It is worth pausing to reflect upon the result in Eq. (9.10). The greater the spacing of phase conductors, the greater the inductance because there is less benefit of magnetic field cancellation from the return current. On the other hand, the greater the conductor radius, the smaller the inductance, because spreading the current across a larger wire cross section means preventing a very strong magnetic field immediately adjacent to where the electric charges are flowing.

Interestingly, an entire circuit can be scaled up by some factor—for example, doubling both the conductor diameter and the spacing—and keep the inductance completely unchanged. At the same time, the resistance of a larger conductor drops dramatically, with the square of diameter. This is why, for large transmission lines, inductive reactance tends to dominate over resistance, whereas for smaller distribution lines, resistance can be equally important.

9.1.4 Geometric Mean Distance and Radius

We will now turn to some of the geometric complexities. Here we write Eq. (9.10) for the per-phase inductance in a more general form, using an equivalent distance D_{eq} and equivalent radius R_{eq} :

$$L_p = 2 \times 10^{-7} \ln \frac{D_{\text{eq}}}{R_{\text{eq}}} \text{ (H/m)} \quad (9.11)$$

First, consider the spacing between phases. The derivation above was shown for a single-phase circuit. What about three phases? Things appear more complicated, since now we have multiple components of return flow to account for, and they are also staggered in time. Fortunately, the enclosed current perspective again comes to our rescue. The easiest case to consider is an equilateral triangle arrangement of three conductors, with a spacing of D between each, as shown in Figure 9.5.

Some transmission line designs approximate that kind of equilateral conductor spacing, but a majority do not. In that case, the symmetry is broken, and the two phases accounting for portions of the return flow will have an unequal effect. The way to analyze this situation is to calculate a representative kind of average for the effective spacing between phases, called the *geometric mean distance* or GMD.

⁹ Many textbooks and engineering tables use a lowercase r' .

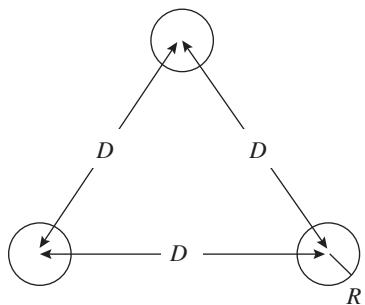


Figure 9.5 Three-phase conductors in an equilateral arrangement, with equal spacing between each.

Clearly, this average won't be exactly right for each individual conductor: for example, in a linear arrangement, if Phase B is in the middle, it will see more field cancellation and would therefore have a smaller flux linkage and inductance than either Phase A or C along the outside. The solution is a practical one: out in the field, the actual phase conductors are physically *transposed* every so many miles, so that each phase gets to take a turn sitting in the middle. Over long distances, we assume that things are approximately even, and each phase comes out with about the same average inductance.¹⁰ This is what is meant when parameters for three-phase lines specify "completely transposed" lines.

The way to calculate the GMD for nonequilateral arrangements applies to any number of phases. We will skip the mathematical derivation here. The key concept is that this GMD provides a meaningful average spacing between the conductors, which takes the place of D and becomes the upper limit of integration D_2 in the expression for flux linkage for a multiphase circuit.

Unlike the arithmetic mean, which takes the sum of n terms and then divides by n , the geometric mean takes the product of n terms and then takes the n th root.¹¹ Here, each term corresponds to the distance between a pair of phase conductors. In practice, three-phase transmission lines are usually described as completely transposed lines with the GMD already calculated.

Example

Consider a three-phase line with linear arrangement and spacing d . The spacing terms between each pair of phases are d , d , and $2d$. The GMD for this arrangement is given by

$$\text{GMD} = D_{\text{eq}} = \sqrt[3]{d \times d \times 2d} = \sqrt[3]{2d^3} \approx 1.26d$$

Another complexity comes with *bundled conductors*, where several wires are held together at intervals with conducting frames to form an individual phase. It is important to note that each part of the bundle is at the same electrical potential; the wires are not insulated from each other. The bundle acts as a single conductor, only with a modified shape. As discussed above in Section 7.2.2, this increases the surface area, promoting heat dissipation and reducing corona losses. We can now appreciate the impact on inductance: spreading out a conductor into a bundle has a similar effect as changing its radius—but without having to increase the cross-sectional area and weight. When

¹⁰ In assuming a balanced three-phase system, there will be some margin of error in any case, since load currents are not precisely balanced either.

¹¹ This way of averaging de-emphasizes very large terms. For example, the arithmetic mean of {1, 10, 100} is 37, while the geometric mean is 10. If three terms represent lengths, which together define a volume, their geometric mean measures the side of a cube of the same volume.

the current is spatially distributed, there will be a less intense magnetic field immediately adjacent to the flowing charges, thus reducing the flux linkage.

The effect becomes more pronounced with increasing number of wires or subconductors in the bundle. It is conceptually helpful to imagine the limit of a very large number of adjacent wires in a circular arrangement, forming a hollow cylindrical conductor. From Ampère's Law, the magnetic field at the surface of the cylinder is the same as it would be for a solid wire at the center carrying the same current. However, the solid wire would additionally have a stronger field closer to it, whereas inside the cylinder the field is zero (since there is no current enclosed). By changing the shape, we are increasing the effective radius of the conductor to the radius of the cylinder, which takes the place of R in the logarithm term to reduce the inductance accordingly. In practice, it is not worth the trouble to approximate a cylindrical shape, as much of the benefit is realized with just a few (two, three, or four) subconductors in the bundle. The effective radius of a bundled conductor for purposes of calculating flux linkage and inductance is the *geometric mean radius* (GMR).

The definition and calculation of geometric mean radius within a bundled conductor is almost but not completely analogous to the geometric mean distance between phases. The caveat is that the GMR accounts not only for the spacing among subconductors within the bundle, but for their individual radii as well. Specifically, the GMR is the geometric mean of all the possible pairs of distances between wires that constitute the bundle, including the distance of each wire to itself—which is simply its effective radius R' . The GMR for a bundle of n subconductors is therefore an average over n^2 terms. More often than not, the subconductors in a bundle will all have the same R' , which condenses the expression a bit, but we must still account for each of them and take the n^2 root. In case a conductor is not bundled ($n = 1$), the GMR simply reduces to R' , so it is never wrong to use the term GMR.

Example

Consider a bundle of four subconductors, each with radius R , in a square arrangement, where the side of the square has length s . To find the GMR, we must take the 16th root of 16 terms, but many of these are repeated. R' occurs four times in the product. The term s occurs twice for each subconductor (since each has two neighbors at that distance), or eight times. The distance between subconductors diagonally across from each other is $\sqrt{2} \times s$, which occurs four times. The GMR for this square bundle is thus given by

$$\text{GMR} = R'_{\text{eq}} = \sqrt[16]{(R')^4 \times s^8 \times (\sqrt{2}s)^4} = \sqrt[16]{(R')^4 \times 4s^{12}}$$

For the sample values of $R = 1$ cm and $s = 6$ cm, this comes out to just under four centimeters:

$$\text{GMR} = R'_{\text{eq}} = \sqrt[16]{0.7788^4 \times 4 \times 6^{12}} \approx 3.93 \text{ cm}$$

For a single cylindrical conductor of the same cross-sectional area as the four-bundle, the equivalent radius would be given by $R_{\text{eq}} = 2R$, with $R'_{\text{eq}} = 0.7788 \times 2R \approx 1.56$ cm. By bundling, the GMR is increased by a factor of $3.93/1.56 \approx 2.5$, substantially reducing the inductance without increasing the amount of material and weight.

To put it all together, we can write a general expression for the per-phase inductance per unit length L_p of a multiphase circuit with bundled conductors as follows:

$$L_p = 2 \times 10^{-7} \ln \frac{D_{\text{eq}}}{R'_{\text{eq}}} \text{ (H/m)} \quad (9.12)$$

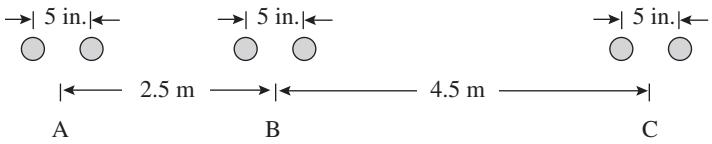


Figure 9.6 Sample three-phase transmission line geometry.

Example

A three-phase transmission line has bundled conductors with two subconductors of 1-in. diameter, spaced 5 in. apart. The phases are in a linear arrangement as shown in Figure 9.6, with separations $d_1 = 2.5$ m and $d_2 = 4.5$ m.

The GMD is given by

$$\text{GMD} = D_{\text{eq}} = \sqrt[3]{2.5 \times 4.5 \times 7} = 4.286 \text{ m}$$

The GMR is given by

$$\text{GMR} = R'_{\text{eq}} = \sqrt[4]{(0.7788 \times 0.5)^2 \times 5^2} = 1.395 \text{ in.} = 0.03544 \text{ m}$$

Note that GMD and GMR need to be in the same units to give the proper ratio. The per-phase inductance per unit length is

$$L_p = 2 \times 10^{-7} \ln \frac{4.286}{0.03544} = 9.6 \times 10^{-7} \text{ H/m}$$

Out of curiosity, let's compare this to the resistance. The per-phase inductive reactance at 60 Hz is

$$X_L = \omega L = 377 \text{ rad/s} \cdot 9.6 \times 10^{-7} \text{ H/m} \approx 3.6 \times 10^{-4} \Omega/\text{m} \approx 0.11 \Omega/1000 \text{ ft}$$

An ACSR conductor of 1-in. diameter is listed with a resistance of about 0.033 Ω per 1000 ft, making 0.016 Ω/1000 ft for the bundle. In this case, the inductive reactance would outweigh the resistance by a factor of about 7 (the X/R ratio).

9.2 Transmission Line Capacitance

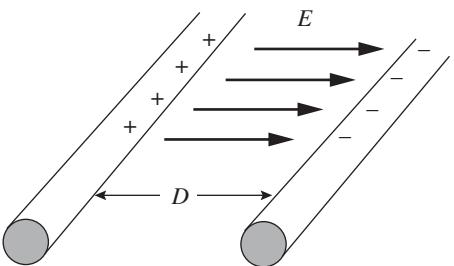
While inductance deals with movement of charge and is mediated by magnetic fields, capacitance is based on the presence of charge *per se*, mediated by electric fields. Recall that the capacitance C of a circuit device is the proportionality constant between stored charge Q and the potential difference V across the device:

$$Q = CV \quad \text{or} \quad C = \frac{Q}{V} \tag{9.13}$$

For a transmission line, we can imagine some charge accumulating on the surface of the conductor, as a result of electrical attraction toward the conductor of opposite phase, or toward the ground, which at any instant resides at a different voltage. This effect will be small compared to the torrential current of charge flowing lengthwise along the conductor. It increases if the opposite phase conductor or ground is close by, supporting a strong electric field that attracts or repels electrons in the direction perpendicular to their flow through the conductor, as illustrated in Figure 9.7. Intuitively, we can imagine capacitance making the conductor surface “sticky” for electric charges.

The physical foundation for deriving the capacitance of a transmission line is Gauss’s law, which—analogous to Ampère’s law for moving charge—relates the electric field across a surface

Figure 9.7 Visualizing capacitance between a pair of conductors.



to the electric charge enclosed by that surface. Gauss's law uses the notion of *electric flux*, which is not to be confused with current; rather, it is an abstract quantity whose spatial density (flux per unit area) defines the electric field. Specifically, the flux density or displacement field D equals the electric field E times the electrical permittivity ϵ (Greek lowercase epsilon), which describes the extent to which the material filling the space is polarized in the presence of the electric field and thus reinforces it. For air, the permittivity of vacuum $\epsilon_0 = 8.854 \times 10^{-12} \text{ F/m}$ is good to about three or four significant figures.¹² Gauss's law can be written as:

$$Q_{\text{encl}} = \oint \epsilon \mathbf{E} \cdot d\mathbf{s} \quad (9.14)$$

This expression captures only the perpendicular component of E , but from symmetry we expect that E should be purely radial, and that it should be the same in every direction.¹³ Again we suppress the length dimension and express all quantities per unit length of a hypothetical conductor.¹⁴ The surface integral becomes a line integral around the conductor, which we are at liberty to consider at any distance r , and we use q for accumulated charge on the conductor per unit length:

$$q = \oint \epsilon \mathbf{E} \times d\mathbf{l} = 2\pi r \epsilon E_r \quad (9.15)$$

or $E(r) = \frac{q}{2\pi r \epsilon}$

The important observation is that this gives rise to a dependence of electric field strength that is inversely proportional to the distance r from the conductor, because it is dividing the enclosed charge by the circumference $2\pi r$ of a hypothetical circle at that distance.

To make a statement about capacitance, we must relate this electric field to a voltage or potential difference between points in space, at various distances from the conductor. Since electric field is the spatial rate of change or gradient of voltage, the voltage difference between two points—which we will choose to lie on a radial line, at distances D_1 and D_2 from the conductor—can be written as the integral of electric field contributions in between:

$$V_{12} = \int_{D_1}^{D_2} E(r) dr \quad (9.16)$$

Substituting our expression from Eq. (9.15) for the electric field as a function of radial distance, we can write

$$V_{12} = \int_{D_1}^{D_2} \frac{q}{2\pi\epsilon} \frac{1}{r} dr = \frac{q}{2\pi\epsilon} \ln \frac{D_2}{D_1} \quad (9.17)$$

¹² Some references use k and k_0 for permittivity.

¹³ The field lines could be pointing outward or inward depending on whether the charge is positive or negative; the sign is not important here.

¹⁴ We also assume that the conductor is long enough that we can safely ignore any effects associated with the geometry of the ends.

Unlike the case of magnetic flux, we need not worry at all about what happens inside the conductor, because it should sustain zero electric field (else it would be a very bad conductor). Therefore, the lower limit of integration is the conductor's own radius, without any adjustment. Equation (9.17) can be directly applied to the voltage difference between two conductors of the same radius R ,¹⁵ separated by a distance D .

There are two situations to consider: either we have two conductors of phases a and b that carry opposite charges, q and $-q$, or one of the conductors is a neutral, approximately at ground potential. For the case of two phase conductors, the opposite charges both contribute to creating the voltage difference between the phase conductors; this accounts for the $2q$ in Eq. (9.18). All else being equal, this would double the voltage. Conversely, a given voltage difference is produced with only half the capacitance if both sides are contributing charge.

$$\begin{aligned} V_{ab} &= \frac{2q}{2\pi\epsilon} \ln \frac{D}{R} = \frac{q}{\pi\epsilon} \ln \frac{D}{R} \quad (\text{volts, line to line}) \\ V_{an} &= \frac{q}{2\pi\epsilon} \ln \frac{D}{R} \quad (\text{volts, line to neutral}) \end{aligned} \quad (9.18)$$

Note that the units in Eq. (9.18) are volts, not volts per meter, because we are looking at the direction perpendicular to the conductor length.

Using the relationship between voltage, charge and capacitance from Eq. (9.13), we solve for capacitance per unit length of conductor, in farads per meter, for opposite phases and for the phase-to-neutral situation:

$$\begin{aligned} C_{ab} &= \frac{q}{V_{ab}} = \frac{\pi\epsilon}{\ln(D/R)} \quad (\text{F/m, line to line}) \\ C_{an} &= \frac{q}{V_{an}} = \frac{2\pi\epsilon}{\ln(D/R)} \quad (\text{F/m, line to neutral}) \end{aligned} \quad (9.19)$$

The notion that the capacitance to neutral is twice that between phases may be a bit counterintuitive, but consistent with the idea of adding the two line-to-neutral capacitances in series to obtain the line-to-line capacitance, as in Figure 9.8. Note that the series addition cuts them in half (since they are akin to admittance, not impedance).

For a three-phase transmission line, the framework is the same as for inductance. We assume the line is completely transposed, so that on average, the capacitance per mile or kilometer is the same for each phase. We use the same geometric mean distance to describe the equivalent conductor spacing D_{eq} between phases as for inductance. For the case of bundled conductors, we use a version of the geometric mean radius that skips the adjustment for R' (which accounted for magnetic flux

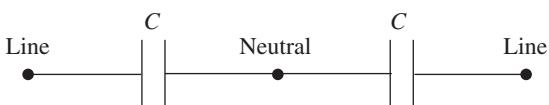
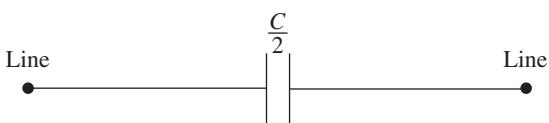


Figure 9.8 Series addition of capacitance.



¹⁵ In the general case, if the radii differ, we would use the geometric mean radius $\sqrt{R_1 R_2}$.

linkage internal to the wire) but is otherwise of the same form, averaging the distance between each possible pair of conductors including the distance to itself (its own radius) to yield an equivalent radius R_{eq} . Again, we assume the phase spacing is much larger than the bundle spacing, and we ignore the irregularities in surface charge that arise from stranded conductors. The capacitance between any one phase and neutral for various line geometries is then given by

$$C_{an} = \frac{2\pi\epsilon}{\ln(D_{eq}/R_{eq})} \quad (9.20)$$

The capacitance for an entire three-phase circuit will be the sum of the three line-to-neutral capacitances.

Example

Let's find the capacitance for the same three-phase line with bundled conductors from the previous example. The GMD is unchanged at

$$D_{eq} = 4.286 \text{ m}$$

but for the GMR we need to eliminate the factor of 0.7788. Recalling that there are two conductors of 0.5-in. radius, 5 in. apart inside their bundle, we have

$$R_{eq} = \sqrt[4]{0.5^2 \times 5^2} = 1.58 \text{ in.} = 0.040 \text{ m}$$

Using these quantities, the per-phase capacitance per unit length comes out to

$$C_{an} = \frac{2\pi\epsilon}{\ln(4.286/0.040)} = \frac{2\pi8.854 \times 10^{-12}}{4.67} = 1.2 \times 10^{-11} \text{ F/m}$$

This capacitance will appear not in series, but in parallel with the line's inductance and reactance, since we are looking at a path directly across the air. Thus, we want to convert it not into a capacitive reactance, but a susceptance, labeled casually as an admittance, at 60 Hz:

$$\begin{aligned} Y_C &= \omega C = 377 \text{ rad/s} \times 1.2 \times 10^{-11} \text{ F/m} = 4.5 \times 10^{-9} \text{ S/m} \\ &\quad = 7.3 \times 10^{-6} \text{ S/mi} \end{aligned}$$

It is worth reflecting on the distinction between capacitance and corona losses. Both phenomena are associated with an accumulation of charge on the conductor surface, and both are represented as a shunt admittance. The effect of the capacitance is to draw a current that is 90° out of phase with the line voltage. This will result in a discrepancy between the line current at the sending and receiving ends, by Kirchhoff's current law and the rules of complex phasor addition. However, the current that appears to vanish into thin air due to capacitance carries away no energy. By contrast, the current associated with corona losses is in phase with the voltage and does irreversible physical work on the air molecules surrounding the transmission line.

9.2.1.1 Ground Effects

Just as the magnetic field beyond the conductor spacing of a single- or multiphase circuit cancels when the enclosed current is zero, it is reasonable to assume that the electric field beyond the conductors cancels because the enclosed net charge is zero. Sometimes these assumptions can be inaccurate, as the physical ground gets involved if the phase currents are not balanced. The extreme case is when ground return is used for the entire line current by design.

The way to calculate inductance and capacitance associated with the ground involves projecting a mirror image of the transmission line conductors on the opposite, subterranean side of the ground surface. This technique is elegant, but we will not cover it here.

Fortunately, the typical situation involves reasonably well balanced circuits. Moreover, for practical and safety reasons, the height above ground is usually quite a bit greater than the spacing between phase conductors. Therefore, the electrical or magnetic interactions between the phase conductors dominate over the interaction with the physical earth, and we can reasonably ignore the ground effects.

9.3 ABCD Parameters

In this section, we will combine the resistive, inductive and capacitive properties of transmission lines derived above into a complete model, which we can use in practice to relate the voltages and currents at either end.

9.3.1 Two-Ports

The standard method for analyzing transmission lines uses a compact matrix notation. The basic model is known as a *two-port*, an extension of the one-port discussed in Section 2.5.1. A two-port is a familiar concept in electronics and telecommunications. It accounts for voltage and current on either end (each “port”) of an abstracted box.

For example, one could make empirical measurements of receiving-end voltages and currents for a set of sending-end voltages and currents, and try to deduce what’s inside the box. Here, we go the other way: based on things we know about transmission lines, we define a two-port to draw ready conclusions about what will be observed externally. Specifically, given the sending-end voltage and current, we can compute the receiving-end quantities, and *vice versa*. We will make one key assumption: namely that the box involves only linear relationships. That is, the sending- and receiving-end voltages and currents are related to each other by a set of linear equations with constant coefficients, consistent with Ohm’s law. To formalize these relationships, we use the transmission matrix:

$$\begin{bmatrix} V_S \\ I_S \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \quad (9.21)$$

which is a shorthand summary of the two equations

$$\begin{aligned} V_S &= A V_R + B I_R \\ I_S &= C V_R + D I_R \end{aligned} \quad (9.22)$$

where the subscript S indicates sending-end and R receiving-end quantities.¹⁶

The idea is that once the four parameters A, B, C, and D are known, then we can solve for the pair of sending- or receiving-end variables given the other pair, meaning that we can predict the effect

¹⁶ There are many variants of the two-port relationships commonly used for analyzing other types of networks. For example, voltages and currents can be paired as independent variables. The sending-end current is generally defined as positive entering the two-port so that positive power corresponds to an *injection*, but the receiving-end current can be defined as either entering or exiting the two-port. We will choose the latter option, where I_R is leaving the two-port and receiving-end power is positive for a *load*. This sign convention has the advantage that we don’t need to switch sign between I_S and I_R ; furthermore, we could string multiple two-ports in series.

of the transmission line under any operating conditions. Most of the remainder of this section will be devoted to the problem of determining these four parameters with a reasonable level of effort or approximation.

But first, let us make some general observations about the symmetry of our model. It stands to reason that it should not matter which end of the transmission line we designate as “sending” or “receiving,” since in practice transmission lines don’t have a directionality assigned to them. A two-port with this property is formally called *reciprocal*. It also should not matter whether power is being injected or consumed at either end, that is, what reference direction we choose for current relative to the sign of voltage. A two-port with this property is called *symmetrical*. With some linear algebra, it can be shown that these conditions require, respectively,

$$\begin{aligned} AD - BC &= 1 \\ A &= D \end{aligned} \tag{9.23}$$

The rationale for the first condition is that, if we switch sending and receiving ends, we would want to use the *inverse* of the transmission matrix:

$$\begin{bmatrix} V_R \\ I_R \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} \begin{bmatrix} V_S \\ I_S \end{bmatrix} \tag{9.24}$$

where

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \frac{1}{(AD-BC)} \begin{bmatrix} D & -B \\ -C & A \end{bmatrix} \tag{9.25}$$

If the *determinant* $(AD-BC) = 1$, it preserves the scaling of all the relationships when we invert them. The rationale for the condition that $A = D$ is more difficult to state in general. However, the fact that $A = D$ will follow directly from applying Kirchhoff’s laws within the specific line models.

The units of the ABCD parameters are worth noting. A and D are dimensionless, per-unit values. B, which multiplies a current to yield a voltage, has units of ohms. C multiplies a voltage to yield a current, so it should be in siemens. Note that all four parameters are, generally, complex quantities—except in special cases, where some of them are purely real or purely imaginary.

9.3.2 Line Models Overview

Every transmission line has *some* resistance, inductance, and capacitance.¹⁷ The key questions are, can any of these attributes be reasonably neglected, and do we need to consider their spatial distribution? Accordingly, there are four common types of transmission line models:

1. The long line model with distributed line parameters, that captures everything we know;
2. The medium line approximation, which simplifies the situation by lumping inductance and capacitance into single representative elements;
3. The short line approximation, which simplifies things further by neglecting shunt capacitance; and
4. The lossless line model, which neglects resistance, although the line may be long.

¹⁷ With the exception of a superconducting line that has no resistance.

An approximate rule of thumb suggests that the medium line model should be used for lines longer than 50 km, and the long line model for 250 km or more when exact answers are required.

All of these models describe a single phase. We assume that a three-phase line would be balanced, with its mutual phase interactions captured within the impedance parameters for the single-phase model, so that the results for voltage and current apply equally to each phase. For power transfer on a three-phase line, we can just multiply by 3. Let's start with the easiest case, the short line.

9.3.3 Short Line Model

The short line approximation neglects the interaction of electric fields surrounding the transmission conductors. This means neglecting all shunt admittance, including discharge through the air by corona losses, and capacitive interaction between phases and between phase and ground. This is a good assumption if the associated currents are small. Adopting the short line model means assuming that the current entering the transmission line equals the current leaving it; there is no way for current to “leak” out of the line.

Note that as shunt admittance varies directly with the length of a line, the shunt or “leakage” current will represent a higher fraction for a longer line carrying the same current. It is thus reasonable to use line length as a criterion for neglecting shunt admittance. For lines shorter than 50 km (30 mi), we would typically expect the shunt current to be a fraction of a percent of the line current, making it safe to ignore for most purposes.

The short line model, shown in Figure 9.9, therefore has only series impedance Z , which includes both resistance R and inductive reactance X_L . For most common transmission lines, the reactance outweighs the resistance, or $X \gg R$. As discussed in Section 9.1, this is a consequence of physical scale: as conductor cross section and spacing are both scaled up proportionally, resistance decreases while reactance stays the same. The larger the dimensions of a three-phase line, the less important its resistance is compared to its inductance. However, line length does not affect the X/R ratio, since series R and X both accrue per unit length.

The relationships between sending- and receiving-end voltages and currents here are so straightforward that transcribing them into ABCD format seems needlessly awkward, but we do it to show the structure and later compare other models.

The first observation is that since the short line model requires that $I_S = I_R$ always, the parameter D (which relates the two currents to each other) must equal 1, and the parameter C (which captures the difference between sending- and receiving-end currents) must equal 0. The second observation is that the sending- and receiving-end voltages will differ exactly by the voltage drop across the series impedance. In other words,

$$V_S = V_R + I_R Z \quad (9.26)$$

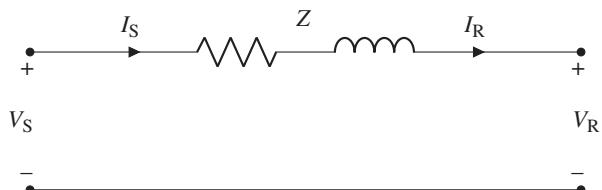


Figure 9.9 Short transmission line model, which accounts for only the series impedance $Z = R + jX_L$. This model requires $I_S = I_R$.

So the series impedance Z is exactly our parameter B in the matrix. In sum, the transmission matrix for the short line model looks like this:

$$\begin{bmatrix} V_S \\ I_S \end{bmatrix} = \begin{bmatrix} 1 & Z \\ 0 & 1 \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \quad (9.27)$$

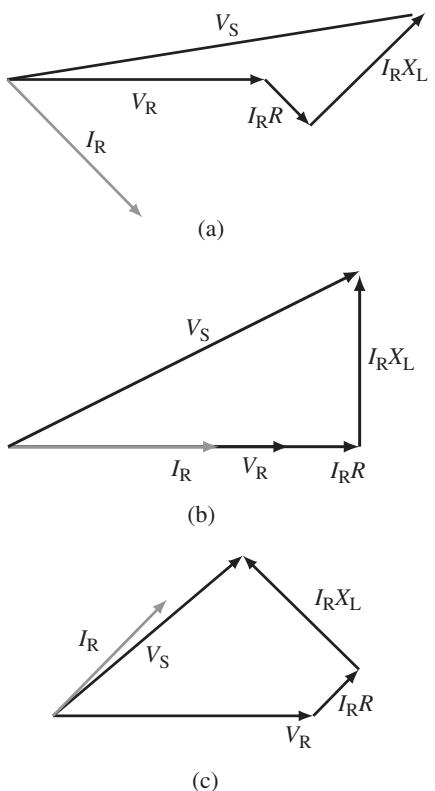
9.3.4 Short Line Phasor Relationship

It is useful to visualize the relationship between V_S and V_R in the complex plane. For convenience, we choose V_R as the reference and let $\angle V_R = 0$. The important observation here concerns the relative direction of the voltage drop, $I_R Z$. Note that the receiving-end current is determined by the load connected there. Based on the (displacement) power factor of that load, this current will either lead or lag the receiving-end voltage by some number of degrees.

The voltage drop $V_S - V_R = I_R Z$ would point in the same direction as I_R if the line were purely resistive. But it has a major component perpendicular to I_R given by $I_R X$, which rotates the voltage drop counterclockwise, since our short transmission line is inductive rather than capacitive. This gives rise to a voltage phase angle difference between the sending and receiving ends. Figure 9.10 illustrates three different load power factor situations, where the transmission line has $X > R$.

Notice that a current in phase with receiving-end voltage—that is, a current associated with supplying real power to the load—will result in a substantial voltage phase angle difference across a primarily inductive line, and not much voltage magnitude difference. As the current phasor rotates away from V_R for a more lagging power factor, the dominant voltage drop component $I_R X$ increasingly points in the same direction as V_R , indicating a substantial voltage magnitude drop between

Figure 9.10 Voltage drop across a short transmission line for a load of lagging, unity, and leading power factor. Note that the magnitude of I on the phasor diagram is arbitrary, but the products $I_R R$ and $I_R X$ have dimensions of voltage and are therefore to scale. (a) Load p.f. = 70% lag, (b) Load p.f. = 100%, and (c) Load p.f. = 70% lead.



sending and receiving ends. This observation is consistent with an important, empirical rule of grid operators: namely that reactive power demand tends to draw down the local grid voltage, much more so than real power, while capacitive loads can produce a voltage rise across an inductive line.¹⁸

The notion of a load creating a voltage rise is so counterintuitive that it warrants a bit more discussion. Doesn't it violate energy conservation, for a load to *increase* the grid voltage at its connection point? In a d.c. setting, this surely wouldn't make any sense. But the key is to remember that we are dealing with complex quantities here. Net energy is only transferred to the extent that current and voltage are aligned with each other in time. Since the capacitive load current is leading the voltage, it entails no net energy consumption. Conversely, the voltage may stretch or shrink in the imaginary direction—out of phase with the current—without affecting net energy transfer.

The short line model is perfectly adequate for developing these qualitative insights. Even some quantitative questions—for example, how much voltage drop to expect across a line for a given load—can often be answered to a good approximation, especially when the line in question is actually short. But if we want to account for the effects of both inductance and capacitance in the transmission line, we need the medium-length model.

9.3.5 Medium Line Model

The medium-length transmission line model augments the series impedance with a shunt admittance. This shunt admittance is almost always dominated by capacitance, and conductance is often ignored entirely (as in Figure 9.11).

One subtlety is the placement of the shunt admittance branch: which side of the series impedance should it go on? Since we must assume that the sending- and receiving-end voltages are different, it follows that the effect of the parallel circuit branch—that is, the current through it—would depend on its placement in the model. There are two ways to “split the difference” for a nice approximation while keeping the components lumped (rather than distributed across the entire line): the T and the π -equivalent circuit.

In the T model (not shown), we split the series impedance into two halves and place the shunt admittance in the middle. That way, we are assigning the average between sending- and receiving-end voltages to the shunt branch. In the π model (shown in Figure 9.11), which is more commonly used in power systems, we split the admittance and place half of it at either end.¹⁹

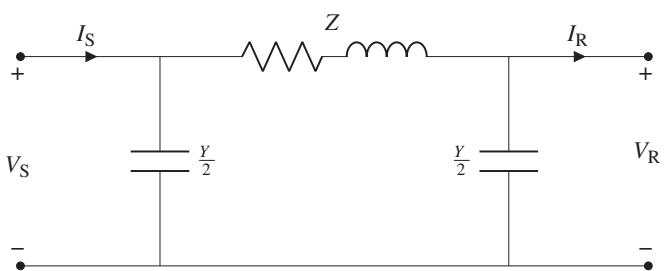


Figure 9.11 Equivalent π -circuit for a medium-length transmission line.

¹⁸ These relationships would be reversed if transmission lines were mostly capacitive rather than inductive in nature, but that is not a common real-world scenario.

¹⁹ The advantage of the π -equivalent is that it fits well with a common network analysis tool, the *bus admittance matrix*, where admittances are assigned to every node in the network (Section 12.4.2). Mathematically, it is easy to allocate $Y/2$ to the end points of a transmission line, whereas it would be hard to account for something in the middle of the line.

Added together, the two “semi”-shunt currents will make a good representation of the total, while the current through the series impedance will be an average between the current entering and the current leaving. We allocate half of the total line admittance to either end. Note that the diagram shows only a shunt capacitance because the shunt conductance is typically considered negligible. Nevertheless, the medium line model allows Y to have a real component.

Deriving the ABCD parameters will take a bit more effort than for the short line. We first write the sending-end voltage as the sum of the receiving-end voltage and the voltage drop. The voltage drop is given by the product of the series impedance and the current through it—but now we must be careful to fully account for this line current as the sum of the receiving-end current I_R , plus the current through the shunt admittance, which is given by $Y/2V_R$. After substituting this expression into the voltage drop, we rearrange terms to express V_S as a linear combination of V_R and I_R , whose coefficients become our A and B parameters:

$$\begin{aligned} V_S &= V_R + Z \left(I_R + \frac{V_R Y}{2} \right) \\ &= \left(1 + \frac{YZ}{2} \right) V_R + ZI_R \end{aligned} \quad (9.28)$$

To get an expression for sending-end current, we write it as the sum of the shunt current at the sending end plus the line current, which itself is I_R plus the shunt current at the receiving end.²⁰ We then substitute Eq. (9.28) for V_S , and finally rearrange terms to express I_S as a linear combination of V_R and I_R whose coefficients will be C and D:

$$\begin{aligned} I_S &= I_R + \frac{V_R Y}{2} + \frac{V_S Y}{2} \\ &= I_R + \frac{V_R Y}{2} + \left[\left(1 + \frac{YZ}{2} \right) V_R + ZI_R \right] \frac{Y}{2} \\ &= Y \left(1 + \frac{YZ}{4} \right) V_R + \left(1 + \frac{YZ}{2} \right) I_R \end{aligned} \quad (9.29)$$

Summarizing the above in matrix form, we can write:

$$\begin{bmatrix} V_S \\ I_S \end{bmatrix} = \begin{bmatrix} \left(1 + \frac{YZ}{2} \right) & Z \\ Y \left(1 + \frac{YZ}{4} \right) & \left(1 + \frac{YZ}{2} \right) \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \quad (9.30)$$

While these ABCD parameters have a certain daunting appearance, it is important to recognize that they are not terribly different from the ones in the short line model; rather, we've introduced a correction that is likely to be quite minor, assuming the shunt admittance does not dominate the transmission line behavior.

Parameters A and D should be fairly close to 1, which will be true as long as YZ is small. The correction will vanish altogether if we neglect admittance (as in the short line model). Another helpful observation is that A and D should be close to purely real, and slightly less than 1. This is because Y is purely imaginary, and Z should be largely imaginary if $X \gg R$, making the product YZ quite real, and negative (since $j^2 = -1$).

Parameter B remains simply the series impedance Z . Parameter C should be a very small number, resembling the default value of zero from the short line model. Again, this certainly holds true for small values of Y . Moreover, we expect C to be almost purely imaginary.

²⁰ Notice that the sum of the two shunt currents at either end is not identical to the current we would get if we just combined the half-capacitances into a single branch because the voltages across them may be different.

A numerical example will illustrate how the medium-length model corrects the short line assumption.

Example

A 200-km long three-phase transmission line is rated 345 kV (line-to-line) and has series impedance per unit length $z = 0.03 + j0.36 \Omega/\text{km}$ and shunt admittance per unit length $y = j4 \times 10^{-6} \text{ S/km}$. Find the sending-end voltage required to deliver 500 MW at unity power factor, if the receiving-end voltage is 95% of rated voltage.

For 200 km length, the impedance and admittance are:

$$Z = 200z = (0.03 + j0.36) \times 200 = 6 + j72 = 72.25 \angle 85.2^\circ \Omega$$

$$Y = 200y = j4 \times 10^{-6} \times 200 = 8 \times 10^{-4} \angle 90^\circ \text{ S}$$

Using the medium line approximation, the ABCD parameters are calculated as follows:

$$\begin{aligned} A = D &= 1 + \frac{YZ}{2} = 1 + \frac{1}{2}(8 \times 10^{-4} \angle 90^\circ \cdot 72.25 \angle 85.2^\circ) \\ &= 1 + 0.0289 \angle 175.2^\circ = 0.97 \angle 0.14^\circ \end{aligned}$$

$$B = Z = 72.25 \angle 85.2^\circ \Omega$$

$$C = Y \left(1 + \frac{YZ}{4} \right) = 8 \times 10^{-4} \angle 90^\circ \left(1 + \frac{0.0289 \angle 175.2^\circ}{2} \right) = 7.88 \times 10^{-4} \angle 90.07^\circ \text{ S}$$

In order to calculate the line current, we first convert the receiving-end voltage into line-to-neutral format:

$$V_{R,LL} = 0.95 \cdot 345 \text{ kV} = 328 \text{ kV}_{LL}$$

$$V_{R,LN} = \frac{1}{\sqrt{3}} 328 \text{ kV} = 189 \text{ kV}_{LN}$$

The line current in each of the three phases is then given by

$$I_R = \frac{1}{3} \frac{500 \text{ MVA}}{189 \text{ kV}} = 0.881 \angle 0^\circ \text{ kA}$$

Note that if the power factor were other than unity, the apparent power in the numerator would be greater than 500 MVA and the current correspondingly greater.

We are now ready to substitute V_R and I_R into the equation for sending-end voltage:

$$V_S = A V_R + B I_R$$

$$V_{S,LN} = 0.97 \angle 0.17^\circ \times 189.2 \angle 0^\circ + 72.25 \angle 85.3^\circ \times 0.881 \angle 0^\circ$$

$$= 199.2 \angle 18.7^\circ \text{ kV}_{LN}$$

Converting back into line-to-line format,

$$V_{S,LL} = 199.2 \times \sqrt{3} = 345 \text{ kV}_{LL}$$

which happens to be exactly the nominal voltage. Expressed in per-unit values (Section 8.7), $V_S = 1.0 \text{ p.u.}$ gives $V_R = 0.95 \text{ p.u.}$ A voltage drop on the order of 5% is plausible for a transmission line operating near its rated capacity (see Section 8.6 on voltage regulation).

9.3.5.1 Charging Current

One consequence of line capacitance is that a finite current flows whenever the line is energized, even under no load. This is called the *charging current*. It is calculated by setting the receiving-end current to zero, and applying the nominal sending-end voltage. Since the line is acting as a capacitive load under this condition, the charging current has a positive (leading) angle. Note that even for a short line, it is necessary to use the medium-length model to account for a charging current.

One could use the ABCD parameters with $I_R = 0$ to solve for I_S in terms of V_S , but that is actually more cumbersome than reviewing the equivalent π -circuit from Figure 9.11 and writing down the relationship by inspection. Note that with $I_R = 0$, the sending-end current I_S has just two components: one through the half-capacitance $\frac{Y}{2}$ at the sending end, and one through the rest of the line, where the impedance Z is in series with the half-capacitance at the receiving end. Thus, we can write:

$$I_S = \frac{V_S Y}{2} + V_S \left(\frac{1}{Z + \frac{Y}{2}} \right) \quad (9.31)$$

Example

The charging current for the transmission line in the previous example, with $V_S = 345 \text{ kV}$, $Z = 72.25\angle 85.2^\circ \Omega$, and $Y = 8 \times 10^{-4}\angle 90^\circ \text{ S}$, can be calculated from Eq. (9.31) as follows:

$$\begin{aligned} I_S &= 345\angle 0^\circ \text{ kV} \cdot 4 \times 10^{-4}\angle 90^\circ \text{ S} + 345\angle 0^\circ \text{ kV} \left(\frac{1}{72.25\angle 85.2^\circ \Omega + \frac{2}{8 \times 10^{-4}\angle 90^\circ \text{ S}}} \right) \\ &= 0.138\angle 90^\circ + 0.142\angle 89.7^\circ \text{ kA} = 280\angle 89.9^\circ \text{ A} \end{aligned}$$

This result is ever so slightly different from what we would have estimated using simply

$$I_S \approx V_S Y = 345\angle 0^\circ \text{ kV} \cdot 8 \times 10^{-4}\angle 90^\circ \text{ S} = 276\angle 90^\circ \text{ A}.$$

9.3.6 Medium Line Qualitative Observations

The corrections in the medium line model allow for two counterintuitive behaviors that result from transmission line capacitance. It makes sense that sending- and receiving-end currents are not the same. Less obvious is the fact that, since $A = D < 1$, I_S could actually be *less* than I_R , and that voltage could actually *rise* rather than drop toward the receiving end, as the capacitive effect from parameter A counteracts the voltage drop ZI_R . Again, we trust that energy conservation will never be violated once we take the complex products of current and voltage. As we saw in Figure 9.10, the power factor of the load connected at the receiving end plays a crucial role in determining the direction of current and voltage drop in the complex plane.

To cement our conceptual understanding, it is helpful to refer back to the phasor diagrams from Figure 9.10 for the short line. Based on our earlier argument that parameter A should be mostly real and slightly less than 1, the basic modification of the diagrams for the medium line is to shorten V_S . The general observation still holds that an inductive load tends to result in a voltage magnitude drop toward the receiving end (requiring a higher V_S to sustain a given V_R), whereas a capacitive load tends to cause a voltage rise. But now, a voltage rise is possible even if the load power factor is near unity or slightly lagging, as long as the current is fairly small. For increasing load current, the voltage drop will dominate, drawing down V_R . This makes physical sense in that the inductive effect of the transmission line (which tends to reduce V_R) depends on current, whereas the capacitive effect (which tends to raise V_R) depends on voltage and is therefore not very sensitive to load.

A related rule known to grid operators is that a highly loaded transmission line will act as an inductive load that “consumes” reactive power, whereas a lightly loaded line will appear as a capacitance in the system that “produces” VARs. Somewhere in the middle is an amount of load for which the transmission line impedance appears neutral. For the special case of the lossless line, we will show that in this situation, $V_R = V_S$, and the load to obtain this condition is called the *surge impedance load* (see Section 7.3.3).

One important practical lesson is that voltage rise on a lightly loaded line can be surprisingly large. This must be taken into account especially when an out-of-service line is initially energized, as it could actually damage equipment.

9.3.7 Long Line Model: Introduction

The most comprehensive and general transmission line model accounts for the fact that voltage and current vary continuously along the length of the line. Although these variations tend to be small, capturing them correctly requires an entirely different kind of mathematical description, using differential equations. We will discover that this rigorous description allows for an entirely new phenomenon: the propagation of waves of currents and voltages.

This raises some interesting conceptual challenges. In power transmission, we generally deal with steady-state quantities, where power is injected or extracted from the network at different locations by virtue of how local voltage and current align. It is impossible to track a package of power being “sent” from generator to load. By contrast, waves on transmission lines are akin to the propagation of *signals*, as in telecommunications, where a wave crest can be seen as traveling in space and time.

While communication signals are intentional, waves on power lines are an unintended artifact of the infrastructure. A key difference is that the *characteristic wavelength* for an a.c. transmission line, which depends on its physical parameters and the a.c. frequency, is very long compared to the actual length of most lines, on the order of thousands of kilometers. By contrast, communications networks deal with signals of much higher frequencies and correspondingly shorter wavelengths, so signal propagation is absolutely central to their analysis. But it is only for very long transmission lines that the waxing and waning of voltage and current due to the wave propagation effects will actually result in appreciable differences between sending and receiving ends, compared to what the medium-length model would predict.

The ABCD parameters we will derive to summarize the results look even more impressive (as they will now be embellished with hyperbolic sines and cosines), but once again will tend to yield only small numerical corrections to the previous model. Readers can skip this section without loss of continuity. The reason to include material on long transmission lines in an introductory course at all is that it highlights some interesting, fundamental methods and insights from physics and engineering. It also illustrates how electric grids produce counterintuitive behaviors merely by virtue of their size.

9.3.8 Long Line Model: Wave Behavior

The problem addressed by the following methodology, illustrated in Figure 9.12, is how to account for the fact that series impedance and shunt admittance are physically spread out along the entire line, while the effects of each—namely, small changes in voltage and current—intermingle with the other. We will use lowercase z and y to denote the distributed line parameters, in ohms per unit length and siemens per unit length (mile or kilometer). Calling the total line length ℓ , we have $Z = z$ and $Y = y$ for the total series impedance and shunt admittance, respectively.

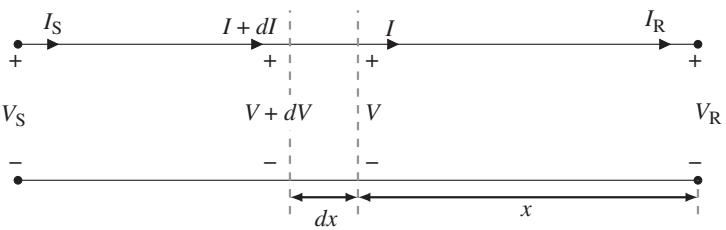


Figure 9.12 Derivation of the interdependence of voltage and current as a function of distance x along a transmission line, with distributed line parameters.

Using the distributed parameters, we can now imagine dividing the shunt admittance into an infinite number of branches, each of which sees an infinitesimally different voltage (instead of just crudely accounting for the different voltages at either end of the line). To keep track of distance along the line, we start with $x = 0$ at the receiving end and let x increase from right to left. We can now write the line voltage as a function of distance, $V(x)$, where $V(0) = V_R$. For every incremental distance Δx that we move along the line, we add a voltage drop ΔV that corresponds to the current multiplied by an incremental impedance $z\Delta x$. In the limit where we shrink these increments to infinitesimal differential elements, we can write the differential equation

$$dV = Iz \, dx \quad \text{or} \quad \frac{dV}{dx} = I(x) z \quad (9.32)$$

which tells us, intuitively enough, that the rate of change of voltage with respect to distance is given by the local current times the impedance per unit length. We amend our notation to make explicit that the current is itself a function of distance, $I(x)$. Specifically, the current will change by an increment ΔI for every additional increment of shunt admittance $y\Delta x$ that is accounted for as we move along the line. In differential form,

$$dI = Vy \, dx \quad \text{or} \quad \frac{dI}{dx} = V(x) y \quad (9.33)$$

where we see that the rate of change of current, or the rate at which current leaks between one spot x on the transmission line and the next, is given by the local voltage times the admittance per unit length.

This pair of differential equations gives us information about the behavior of both $V(x)$ and $I(x)$, but first we must disentangle the two variables. This is done by differentiating Eqs. (9.32) and (9.33) again with respect to x , which cleverly allows us to substitute I and V back in across the two equations:

$$\frac{d^2V}{dx^2} = z \frac{dI}{dx} \quad \text{and} \quad \frac{d^2I}{dx^2} = y \frac{dV}{dx} \quad (9.34)$$

$$\frac{d^2V}{dx^2} = yz \, V(x) \quad \text{and} \quad \frac{d^2I}{dx^2} = yz \, I(x) \quad (9.35)$$

We now have a matching pair of second-order differential equations for voltage and current as functions of distance along a transmission line. Their format is very familiar to those who have studied oscillators or wave behavior in physics: each equation states that the second derivative (i.e., the rate of change of the rate of change, or acceleration) is directly related to the quantity itself (voltage or current, respectively) at any given location x , with a proportionality constant yz .

Perceptive readers may note that in order to see oscillatory behaviors, we really need a negative sign in these equations: we don't want the quantity to *increase* when it is already large, but instead

tend to return to its neutral value. This important minus sign arises from y and z each having imaginary parts. However, y and z may also have real parts, which will keep things interesting.

A differential equation is like a job announcement with a list of qualifications for a mathematical function. We happen to know just what type of function meets the above requirements: an exponential. When writing down $V(x)$ and $I(x)$, it's important to consider all possible ways the criteria could be met. The general solution is the sum of a positive and a negative exponential, with coefficients C_1 and C_2 to be determined later from information about the boundary conditions.

Since the constant factor yz comes from differentiating twice, the factor in the exponent is \sqrt{yz} . This quantity is so important that we give it a special name, the *propagation constant* γ (Greek lowercase gamma). Note that γ has dimensions of inverse length, which appropriately yields a dimensionless exponent γx . Using

$$\gamma \equiv \sqrt{yz} \quad (9.36)$$

we can write for voltage

$$V(x) = C_1 e^{\gamma x} + C_2 e^{-\gamma x} \quad (9.37)$$

and verify that the solution meets the requirements of the differential equation:

$$\frac{d^2 V}{dx^2} = \gamma^2 (C_1 e^{\gamma x} + C_2 e^{-\gamma x}) = \gamma^2 V(x) = yz V(x) \quad (9.38)$$

Since voltage and current are interrelated, we use Eq. (9.32) to write an expression for current:

$$I(x) = \frac{1}{z} \frac{dV}{dx} = \frac{\gamma}{z} C_1 e^{\gamma x} - \frac{\gamma}{z} C_2 e^{-\gamma x} \quad (9.39)$$

Let us define another useful quantity, the *characteristic impedance* Z_C of the transmission line:

$$Z_C \equiv \sqrt{\frac{z}{y}} \quad (9.40)$$

Note that with length units cancelling, z/y has dimensions of impedance squared, correctly giving Z_C in ohms. Z_C not only simplifies the notation, but is a useful benchmark to gauge the effect of load impedance connected to the line. It is also called the *surge impedance*.

We can now work out C_1 and C_2 in terms of other variables we know, namely the receiving-end voltage and current values where $x = 0$. We have two equations with two unknowns:

$$\begin{aligned} V(0) &= V_R = C_1 + C_2 \\ I(0) &= I_R = \frac{1}{Z_C} (C_1 - C_2) \end{aligned} \quad (9.41)$$

Solving and substituting C_1 and C_2 back into the equations for voltage and current, we obtain:

$$\begin{aligned} V(x) &= \frac{1}{2}(V_R + I_R Z_C) e^{\gamma x} + \frac{1}{2}(V_R - I_R Z_C) e^{-\gamma x} \\ I(x) &= \frac{1}{2Z_C}(V_R + I_R Z_C) e^{\gamma x} - \frac{1}{2Z_C}(V_R - I_R Z_C) e^{-\gamma x} \end{aligned} \quad (9.42)$$

Different aspects of physical behavior are captured by the positive and negative, real and imaginary parts of the exponential term, respectively. To clarify, we break up the propagation constant into its real and imaginary parts:

$$\sqrt{yz} = \gamma = \alpha + j\beta \quad (9.43)$$

which allows us to separate the exponential terms in Eqs. (9.42). Note that we expect α to be small, assuming y and z are both mostly positive and imaginary, which makes their product negative and real, and its square root γ positive and imaginary.

$$\begin{aligned} V(x) &= \frac{1}{2}(V_R + I_R Z_C) e^{\alpha x} e^{j\beta x} + \frac{1}{2}(V_R - I_R Z_C) e^{-\alpha x} e^{-j\beta x} \\ I(x) &= \frac{1}{2Z_C}(V_R + I_R Z_C) e^{\alpha x} e^{j\beta x} - \frac{1}{2Z_C}(V_R - I_R Z_C) e^{-\alpha x} e^{-j\beta x} \end{aligned} \quad (9.44)$$

The positive real exponent αx is eye-catching, as it suggests the possibility of voltages and currents exploding into large values over long distances (albeit slowly, since α should be small). The negative real exponent $-\alpha x$ appears more benign, and more obviously associated with a decay phenomenon that gives α the name *attenuation constant*. Meanwhile, the imaginary exponent $j\beta x$ produces a counterclockwise rotation in the complex plane, clockwise for $-j\beta x$. Because this changes the relative phase angle as a function of distance, β is called the *phase constant*.²¹ Keep in mind that $V(x)$ and $I(x)$ are phasor quantities that also oscillate in time, and that only their real parts are actually observable.

The physical interpretation is that we are looking at a pair of traveling waves, propagating along the transmission line. Let us revisit our choice of coordinate system, where we set $x = 0$ at the receiving end. This convention stems from the perspective of loads being controlled by the customer at the receiving end, while power engineers figure out what they need to do at the sending end in order to maintain receiving-end voltage within an acceptable range. But physically, transmission lines appear symmetrical, regardless of which end sends or receives power. There is no reason a wave signal should propagate in one direction and not the other. In fact, there is another pair of equations with an equal claim to representing a true phenomenon, in which the sign of the distance coordinate x is reversed. This phenomenon is called the *reflected wave*, and its effects are added to the *incident wave* represented by the original equations.

A propagating wave will be reflected from any boundary where the impedance changes—that is, where the line terminates and a load is connected. This is analogous to light being reflected off a surface with a different refractive index. In the case of a single impulse disturbance (e.g., a lightning strike), we could watch the signal travel from its physical origin down the line, and literally bounce back from the boundary. How much of the signal gets reflected, and how much continues on its way with diminished amplitude, depends on the relationship between the line's characteristic impedance and the load impedance on the other side of the boundary. When the two impedances are equal, power transfer across the boundary is maximized; thus the term *impedance matching*.²² From the point of view of the sending end, the load simply “blends in” and the line appears infinite. (After all, how is an electron to know where the power line ends and the load begins?) In Eqs. (9.44), we can see that if the load impedance, which must be V_R/I_R by Ohm's law, equals Z_C , the entire term with the negative exponential vanishes.

For the steady-state condition of continuous a.c. power transfer, the wave reflection is difficult to visualize. Its result, though, is simply that the voltage and current phasors observed at any point x along the line will be the sum of the incident and reflected waves. The crucial insight here is that the phase rotation due to β will cause cancellation, even if it seems that we are adding positive components. Specifically, if the reflected wave has the same magnitude but is 180° out of phase

²¹ The units here are just inverse length for α , and radians per unit length for β . Engineering fields that deal more routinely with signal attenuation use logarithmic units of *nepers* per length for α .

²² This is why loudspeakers, for example, show an impedance in Ω . When this is not matched with the input, the music sounds faint.

with the incident wave, the physically measured result is zero. This is the case for current if the line is terminated by an open circuit, and voltage at the point of a short circuit.

Propagating waves have a wavelength that relates directly to the frequency (determined by the signal) and the propagation speed (determined by the medium, in this case the characteristic impedance). We will address these relationships in the section on lossless lines, where γ is purely imaginary and the algebra is much cleaner. Roughly speaking, the wavelength λ for typical transmission lines at 50 or 60 Hz is on the order of several thousand kilometers, which would correspond to two complete reversals of voltage and current direction. Only a small fraction of such a phase rotation is actually sustainable for steady-state power transfer, and those limits are typically encountered for a.c. lines of several 100 km.

By way of qualitative summary, we could say that long transmission lines act like they have a mind of their own: they certainly behave as significant entities within a circuit, rather than just passive connectors between generators and loads. We have shown that voltages and currents can both increase and decrease along the length of a line, merely due to the interaction of the line's own series impedance and shunt admittance along the way. While it is theoretically interesting to be able to determine voltages and currents for any position x , we may be concerned in practice with just the values at the receiving end, V_R and I_R . This brings us back to the two-port model formulation.

9.3.9 Long Line Model: ABCD Parameters

In order to capture the effects of propagating wave behaviors with ABCD parameters, we rearrange Eqs. (9.42) so as to separate terms with V_R from those with I_R , instead of separating them by positive and negative exponentials. For this we use *hyperbolic* sine and cosine functions, which concisely represent sums and differences of these exponential terms. Specifically, using the equalities

$$\sinh \gamma x = \frac{1}{2}(e^{\gamma x} - e^{-\gamma x}) \quad \text{and} \quad \cosh \gamma x = \frac{1}{2}(e^{\gamma x} + e^{-\gamma x}) \quad (9.45)$$

we reformat Eqs. (9.42) into

$$\begin{aligned} V(x) &= \cosh \gamma x V_R + Z_C \sinh \gamma x I_R \\ I(x) &= \frac{1}{Z_C} \sinh \gamma x V_R + \cosh \gamma x I_R \end{aligned} \quad (9.46)$$

If we now let x equal the total length l of the line, we have produced expressions for the sending-end quantities, along with specific ABCD parameters. In matrix form:

$$\begin{bmatrix} V_S \\ I_S \end{bmatrix} = \begin{bmatrix} \cosh \gamma l & Z_C \sinh \gamma l \\ \frac{1}{Z_C} \sinh \gamma l & \cosh \gamma l \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \quad (9.47)$$

Example

Consider a transmission line of length $l = 200$ mi with the following impedance values:

$$z = 0.150 + j 0.600 = 0.618 \angle 75.96^\circ \Omega/\text{mi} \quad Z = 123.6 \angle 75.96^\circ \Omega$$

$$y = 6.00 \times 10^{-6} \angle 90.00^\circ \text{ S/mi} \quad Y = 12.00 \times 10^{-4} \angle 90.00^\circ \text{ S}$$

For the characteristic impedance Z_C , we find

$$Z_C = \sqrt{\frac{z}{y}} = \left(\frac{0.618 \angle 75.96^\circ}{6 \times 10^{-6} \angle 90.00^\circ} \right)^{\frac{1}{2}} = (1.03 \times 10^5 \angle -14.04^\circ)^{\frac{1}{2}} = 321 \angle -7.02^\circ \Omega$$

illustrating that Z_C represents an entirely different quantity than Z , even though they are both in ohms. The propagation constant γ comes out to

$$\gamma l = \sqrt{yz} = (0.618\angle 75.96^\circ \times 6 \times 10^{-6} \angle 90.00^\circ)^{\frac{1}{2}} = 0.00193\angle 82.98^\circ \text{ mi}^{-1}$$

In rectangular coordinates,

$$\gamma = \alpha + j\beta = 0.000736 + j0.00192 \text{ (mi}^{-1}\text{)}$$

For the entire line,

$$\gamma l = 0.00193\angle 82.98^\circ \text{ mi}^{-1} \times 200 \text{ mi} = 0.386\angle 82.98^\circ = 0.0472 + j0.383$$

and evaluating the hyperbolic trigonometric functions,²³ we get

$$\sinh \gamma l = 0.044 + j0.374 = 0.376\angle 83.29^\circ$$

$$\cosh \gamma l = 0.929 + j0.018 = 0.958\angle 1.11^\circ$$

This produces the following ABCD parameters:

$$A = D = \cosh \gamma l = 0.958\angle 1.11^\circ$$

$$B = Z_C \sinh \gamma l = 321\angle -7.02^\circ \times 0.376\angle 83.29^\circ = 121\angle 76.27^\circ$$

$$C = \frac{1}{Z_C} \sinh \gamma l = \frac{0.376\angle 83.29^\circ}{321\angle -7.02^\circ} = 0.00117\angle 90.31^\circ$$

The results meet expectations: A and D should be close to 1, B should be close to the series impedance Z , and C should be small. The results are also quite similar to the medium line approximation, which yields:

$$A = D = 1 + \frac{YZ}{2} = 0.928\angle 1.11^\circ$$

$$B = Z = 123.6\angle 75.96^\circ$$

$$C = Y \left(1 + \frac{YZ}{4} \right) = 0.00116\angle 90.53^\circ$$

We would get a more significant discrepancy between the medium and long line model by increasing either γ or l . The best choice of model will depend on the nature and length of the line, and very much on the accuracy expected from the particular calculation.

9.3.9.1 Lumped-Circuit Equivalent

The results from the long line model can be converted to the format for the medium line model. This will give the exact relationships between sending- and receiving-end quantities, but allow us to draw a conventional π -circuit diagram with lumped rather than distributed parameters. To do this, we determine an equivalent Y' and Z' which, when used in place of Y and Z in the medium-line formulation, will provide the same ABCD parameter values as the long line model. The relationships are:

$$Z' = Z_C \sinh \gamma l = \sqrt{\frac{Z}{Y}} \sinh \gamma l \quad (9.48)$$

and, after some algebra that we will skip here,

$$\frac{Y'}{2} = \frac{Y}{2} \frac{\tanh(\gamma l/2)}{\gamma l/2} \quad (9.49)$$

²³ Not every calculator offers sinh and cosh of complex values, but the internet obliges.

The shorter the line and the smaller the propagation constant, the closer the correction factors relative to the medium length model will be to 1. Note that we can equally well use Y and Z instead of distributed parameters y and z in order to determine γ and Z_C .

9.3.10 Lossless Line

It is sometimes appropriate to neglect resistance and treat transmission lines as *lossless*, where the both series impedance Z and shunt admittance Y are entirely imaginary, with resistance $R = 0$ and conductance $G = 0$. In the short and medium line model, it is straightforward to see how this simplifies the arithmetic a bit. This section focuses on simplifications to the long line model resulting from the assumption that the line is lossless.

First, note that the propagation constant $\gamma = \alpha + j\beta$ will have only an imaginary part β . There will be no attenuation, meaning that $I(x)$ and $V(x)$ will maintain the same magnitude along the length of the line, but rotate according to the $e^{j\beta x}$ term in Eq. (9.44).

The wave propagation properties are much tidier to describe for the lossless than the general case. Since the phasor rotation with respect to distance occurs at a rate of β radians per mile, and 2π radians correspond to a complete cycle, we can readily solve for the wavelength λ and propagation velocity v :

$$\lambda = \frac{2\pi}{\beta} \quad \text{and} \quad v = \lambda f = \frac{2\pi f}{\beta} \quad (9.50)$$

Since $\beta = -j\sqrt{zy}$, we can use $z = j\omega L$ and $y = j\omega C$ to alternatively write β in terms of inductance L and capacitance C per unit length²⁴

$$\beta = \omega \sqrt{LC} \quad (9.51)$$

and substitute $\omega = 2\pi f$ to give

$$\lambda = \frac{1}{f \sqrt{LC}} \text{ mi} \quad \text{and} \quad v = \frac{1}{\sqrt{LC}} \text{ mi/s} \quad (9.52)$$

The value of $\beta \approx 0.002$ per mile from our earlier example would give a wavelength $\lambda \approx 3000$ mi and, at 60 Hz, a propagation speed $v \approx 180,000$ mi/s, or 0.97 the speed of light.

When a line is terminated in a load that matches its *characteristic impedance*, this is called the *surge impedance load* (SIL, see Section 7.3.3). The characteristic impedance Z_C for the general case is given by the ratio of series impedance to shunt admittance

$$Z_C = \sqrt{\frac{R + j\omega L}{G + j\omega C}} \quad (9.53)$$

where quantities may be given per unit length, but length cancels.

In the lossless case, where series resistance R and shunt conductance G are zero, Z_C is purely real, and simplifies to

$$Z_C = \sqrt{\frac{L}{C}} \quad (9.54)$$

As noted earlier, when γ is purely imaginary, the magnitude of voltage and current along the length of the line are unchanged; only their angle in the complex plane rotates. The surge impedance load

²⁴ Some texts use L and C for the entire line, in which case we must explicitly divide by the length l .

in watts or volt-amperes is given by

$$\text{SIL} = \frac{V^2}{Z_C} \quad (9.55)$$

As introduced in Section 7.3.3, this metric provides a useful reference in practice, even approximately for a line that is not exactly lossless. If the line's characteristic impedance and receiving-end voltage are known, it is easy to compare the connected load to the SIL. If the load is greater than SIL, the receiving-end voltage will be less than sending end; the line's inductive properties dominate and the line will "consume" VARs. If the load is less than SIL, there will be a voltage rise; the line's capacitive properties dominate and it will "generate" VARs.

The ABCD parameters for a lossless long line appear simpler than those for the long line, because the hyperbolic sine and cosine terms for purely imaginary arguments simplify²⁵ to make parameters A and D purely real and B and C purely imaginary:

$$A(x) = D(x) = \cos \beta x$$

$$B(x) = jZ_C \sin \beta x$$

$$C(x) = \frac{j}{Z_C} \sin \beta x$$

$$\begin{bmatrix} V_S \\ I_S \end{bmatrix} = \begin{bmatrix} \cos \beta x & jZ_C \sin \beta x \\ \frac{j}{Z_C} \sin \beta x & \cos \beta x \end{bmatrix} \begin{bmatrix} V_R \\ I_R \end{bmatrix} \quad (9.56)$$

Problems and Questions

9.1 Refer to Table 7.1.

- (a) Check how the inductance per meter and the inductive reactance per mile are consistent with the other information listed for the 138-kV line.
- (b) Check how the capacitance per meter and the capacitive reactance per mile are consistent with the other information listed for the 138-kV line. Also estimate the shunt admittance in siemens per mile.

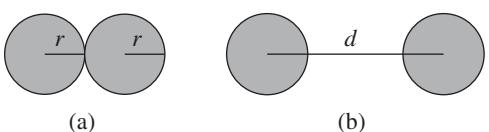
9.2 Look up the specifications of a "Partridge" conductor online. Suppose a single-phase transmission line is made of (nonbundled) Partridge conductors, spaced 10 m apart. Find the inductance and inductive reactance per mile, at 60 Hz.

9.3 Refer to Figure 9.13 showing the cross section for two different conductors.

- (a) Find the geometric mean radius (GMR) for the conductor on the left, composed of two immediately adjacent strands, each with radius r .

Figure 9.13 (a) Adjacent conductor strands.

(b) Conductor strands separated by a distance d within a bundled conductor.



²⁵ $\cosh j\beta x = \cos \beta x$ and $\sinh j\beta x = \sin \beta x$.

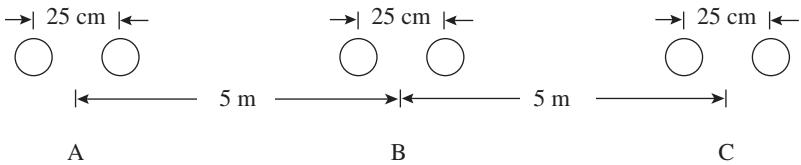


Figure 9.14 Cross section of a three-phase transmission line with bundled conductors.

- (b) Suppose you separate the two strands, effectively turning your stranded into a bundled conductor, as shown on the right. What happens to GMR of the two-conductor bundle as the distance d is increased, and how does this affect the inductance?
 - (c) What can you say about the ratio of reactance to resistance (X/R) for the stranded versus the bundled conductor?
- 9.4** The bundled conductors in Figure 9.14 have a physical radius of 0.75 cm for each subconductor. The line is completely transposed.
- Find the series inductance per phase in mH/km and the inductive reactance per phase in Ω/km at 60 Hz.
 - Find the shunt capacitance per phase in $\mu\text{F}/\text{km}$ and the admittance per phase in $\mu\text{S}/\text{km}$ at 60 Hz.
- 9.5** Look up the specifications for “Flamingo” and “Bluebird” ACSR conductors. Compare the per-phase resistance and inductive reactance per mile of a transposed three-phase transmission line at 60 Hz, with a flat linear phase spacing of 6 m, using (a) a single Bluebird conductor, or (b) a bundle of three Flamingo subconductors spaced 20 cm apart in an equilateral triangle arrangement. Also compare the cross-sectional areas and tabulated current carrying capacities, and comment on the benefits of bundled conductors.
- 9.6** A 25-km, 34.5-kV, 60-Hz three-phase line has a per-phase series impedance $z = 0.19 + j0.34\Omega/\text{km}$. The load at the receiving end absorbs 10 MVA at 33 kV
- Find the ABCD parameters using the short line approximation.
 - Find the sending-end voltage for a load power factor of 0.9 lagging.
 - Find the sending-end voltage for a load power factor of 0.9 leading.
 - Discuss the significance of the load power factor for voltage drop.
- 9.7** Consider a medium-length line with per-phase series impedance $Z = 70.3\angle84.8^\circ\Omega$ and shunt admittance $Y = 8.4 \times 10^{-4}\angle90^\circ\text{ S}$.
- Find the ABCD parameters for this line.
 - Suppose the line-to-neutral receiving-end voltage for this line is $V_R = 220\text{ kV}$, and the per-phase receiving-end current is $I_R = 500\text{ A}$ at unity power factor. Find the sending-end voltage and sending-end current.
 - Repeat with the same receiving-end current magnitude but at a power factor of 0.9 lagging, and comment on the result.
- 9.8** Find the efficiency of the transmission line in the previous problems, based on the I^2R line losses and real power delivered to the load.

- 9.9** Consider a medium-length line with series impedance $Z = 80\angle 80^\circ \Omega$, shunt admittance $Y = 0.0007\angle 90^\circ \text{ S}$, and a sending-end voltage of 230 kV.
- Roughly estimate the charging current for this line.
 - Properly calculate the charging current.
 - Explain the difference between the answers from (a) and (b).
 - Find the receiving-end voltage V_R at no load, when $V_S = 230 \text{ kV}$.
- 9.10** A three-phase, 60-Hz, 345-kV transmission line of 150 km length has a resistance of $0.036 \Omega/\text{km}$, inductance 0.8 mH/km , and capacitance $0.011 \mu\text{F/km}$.
- Find the series impedance and the shunt admittance of the line.
 - Find the appropriate ABCD parameters.
 - Find the sending-end voltage for a three-phase load of 270 MVA at a power factor of 0.9 lagging, if the receiving-end (line-to-line) voltage is 330 kV.
 - Find the (per-phase) sending-end current for the same situation.
- 9.11** A 500-km, 500-kV, 60-Hz line has per-phase series impedance $z = 0.03 + j0.35 \Omega/\text{km}$ and shunt admittance $y = j4.4 \times 10^{-6} \text{ S/km}$.
- Find the characteristic impedance Z_C .
 - Calculate the product of propagation constant and length.
 - Find the exact ABCD parameters for this line.
 - Find the ABCD parameters using the medium line approximation, and comment.
- 9.12** Consider a transmission line with distributed parameters $z = j0.600\Omega/\text{mi}$ and $y = j6.00 \times 10^{-6} \text{ S/mi}$ at 60 Hz.
- Find the wavelength λ of signal propagation.
 - If the line-to-neutral voltage is 200 kV, at what per-phase load would you expect a flat voltage magnitude profile along the length of this transmission line?
 - Explain in your own words why you don't need to know the length of the line to answer the above questions.
- 9.13** Assume the line in the previous problem retains its physical shape with the same inductance and capacitance per mile, but is now operated at an a.c. frequency of 400 Hz.
- What are the new values of γ and λ at this frequency?
 - Qualitatively, explain how you would expect the line to behave differently at the higher frequency.

10

Machines

Electric generators and motors—collectively referred to as *machines*—are devices designed to take advantage of *electromagnetic induction* in order to convert movement into electricity, or *vice versa*. The phenomenon of induction (introduced in Section 1.5.4) can be summarized as follows: an electric charge, in the presence of a magnetic field in relative motion to it—either by displacement or changing intensity—experiences a force in a direction perpendicular to both the direction of relative motion and of the magnetic field lines. Acting on the many charges contained in a conducting material—usually, electrons in a wire—this force becomes an *electromotive force (emf)* that produces a voltage or potential drop along the wire and thus causes an electric current (the *induced current*) to flow.

A generator is designed to obtain an induced current in a conductor (or set of conductors) as a result of mechanical movement, which is utilized to continually change a magnetic field near the conductor. The generator thus achieves a conversion of one physical form of energy into another—energy of motion into electrical energy—mediated by the magnetic field or *flux* (Section 2.6) that exerts forces on the electric charges. In this way, a generator is the opposite of an electric motor, which accomplishes just the reverse: the motor converts electrical energy into mechanical energy of motion, likewise mediated by the magnetic field.

The physical principles for electric generators and motors are exactly the same. In fact, just about any actual generator can be operated as a motor, and *vice versa*.¹ To achieve the best possible performance, though, there are many subtleties of design that optimize a given machine for one or the other task. These subtleties have to do largely with geometry, although the choice of materials may also be important. Indeed, geometry is what distinguishes the many different types of specialized generators: the particular way in which the conducting wires are arranged within the generator determines the spatial configuration of the magnetic flux, which determines the behavior of the machine under various circumstances.

There are almost innumerable variations on generator and motor design, spanning many orders of magnitude in size and serving any imaginable purpose. Remarkably, nearly two centuries after the invention of the first direct-current motors, development continues to advance—today, with a fusion of old-school magnetic principles and modern power electronics, and with electric vehicles as a prime application space. A comprehensive discussion of the many different types of electrical machines would far exceed the scope of this text. This chapter instead concentrates on the basic principles of operation, which apply equally to motors and generators, and on the most common type of generator used in legacy power systems, the synchronous generator.

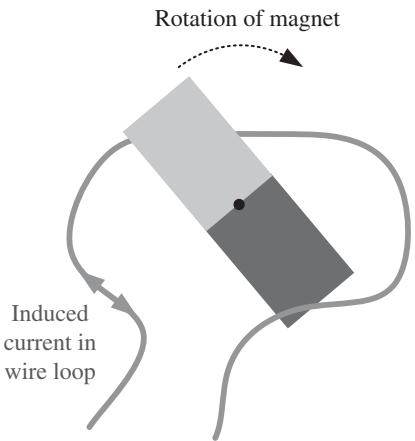
¹ Examples of using single machines for both generating and motoring include pumped hydroelectric storage plants, where large water pumps are operated reversibly as turbine generators.

We begin with a stylized, rudimentary “simple generator” to establish the relationship between magnetic flux and mechanical force, and give just a cursory overview of different types of machines based on this understanding. To emphasize tactile learning, we describe how the reader can actually build a simple machine, the “paper clip motor” that runs on a battery. Second, we describe at some length the standard type of generator used in legacy power systems, the *synchronous machine*. Rather than going into the details of its construction and the many subtle variations among specific designs, we focus on the operation of generators in the system context, emphasizing how synchronous generators can control variables like voltage, frequency, real, and reactive power. Here we try to build some intuition about the interaction between generators, which is fundamental to the overall performance and stability of an alternating current (a.c.) system. We then discuss the *induction machine*, which remains the most common motor load today. Up to this point, the description is entirely qualitative and avoids mathematical abstraction. The chapter concludes with a section on modeling generators as equivalent electrical circuits, which draws on the tools developed in Chapters 8 and 9, but skips many details (such as the ramifications of specific design choices within any given class of machines). The goal is to develop a conceptual understanding of the workings of an electric generator, including its operating constraints and limitations, so as to appreciate its function from the perspective of the power system as a whole. This chapter also ignores the *prime mover*, or whatever energy source pushes the turbine, which is addressed in Chapter 15.

For those already familiar with transformers, it is worth contemplating the similarities. In both cases, generator or transformer, we are dealing with two sets of windings, separated electrically but coupled magnetically. And in both, we are transferring energy to the secondary winding by way of a magnetic flux linkage. In the case of the transformer, the energy comes from some power source attached to the primary winding, which is doing work on electrons to force an alternating voltage and current inside the conductor. In the case of the synchronous generator, the primary winding is replaced by the rotor winding with a direct current that produces a magnetic field. The work is done mechanically by forcing that constant magnetic field to rotate relative to the secondary winding. It’s as if we broke a transformer apart, and instead of forcing the current to change direction inside the primary winding, we just push the whole copper coil around in space relative to the secondary winding.

10.1 The Simple Generator

For the purpose of developing a conceptual understanding of a generator, let us begin by considering a greatly simplified setup that includes only a single wire and a bar magnet. The objective is to induce a current in the wire; in a real generator, this wire corresponds to the *armature*, or the conductors that are electrically connected to the load. This can be accomplished by moving the wire relative to the magnet, or the magnet relative to the wire, so that the magnetic field at the location of the wire increases, decreases, or changes direction. In order to maximize the wire’s exposure to the magnet, since the field decreases rapidly with distance, we form a *loop* of wire that surrounds the bar (with just enough space in between). Now we can produce an ongoing relative motion by rotation: either we rotate the magnet inside the loop, or we hold the magnet fixed and rotate the loop around it. The analysis of the magnetic field is essentially the same in either case, but since the first type of arrangement turns out to be generally more practical, we shall choose it for our example (see Figure 10.1). We now have a bar magnet spinning around inside a loop of wire. How do we analyze the effect of this rotating magnetic field in terms of induced current? It might seem

Figure 10.1 The notional “simple generator.”

as though this would require some tedious analysis of magnetic field lines, including their directions with respect to the various sections of wire, and so forth. Fortunately, it is possible to use a comparatively simple approach. The key is to apply the notion of *magnetic flux* (see Sections 1.5.3 and 2.6), which equals the magnetic field multiplied by the area it crosses. Stated differently, the magnetic field represents the density of the flux in space. The flux is a convenient quantity in this context because the total flux emanating from the bar magnet does not change. The flux is also more easily related to the induced current, especially in a loop of wire. It turns out that the induced current in a closed loop is directly proportional to the flux *linking* (i.e., going through) this loop, irrespective of the spatial distribution of the flux inside the enclosed area. This means that there is no need to consider individually every point of the wire and its particular relationship to a magnetic field somewhere; rather, the entire loop and all the flux through it can be treated as compact quantities, joined in a straightforward relationship: the amount of current induced in the loop of wire will be directly proportional to the rate of change of the magnetic flux linking it.

How, then, can we specify the rate of change of the flux over time as our bar magnet spins? In actuality, the shape of the magnet would come into play, accounting for whatever particular bending of the magnetic field lines occurs, especially at the edges of the poles. For the sake of this example, however, we imagine some sort of ideal magnet that produces a uniform flux across the entire area enclosed by the loop. The *flux linkage* of this loop then varies according to the intersected area, which is in turn described by a *sinusoidal* mathematical function (see Section 3.1). When the magnet points in a direction parallel to the plane of the loop, none of its flux intersects the loop; we would say that the flux linkage of the loop at that time is zero. As the magnet rotates, more and more of its flux goes through the loop, reaching a maximum when the magnet is perpendicular to the plane of the loop. As it continues to rotate, the flux linkage decreases and becomes zero again. Then the magnet points across the loop in the opposite direction; we would now say that the flux linkage is negative. With continued rotation, the flux over time could be plotted out as an oscillating function, resembling more or less closely the mathematical sine or cosine wave. This progression is illustrated in Figure 10.2.

The induced voltage or *emf* and current in the wire are in turn given by the rate of change of the flux.² Technically, we should say the *negative* rate of change, though this sign convention is only to

² Assuming that the load connected to the generator is purely resistive (see Section 3.2), the voltage and current in the wire will be exactly proportional to each other at all times. They can thus be used interchangeably in the following discussion, and, for simplicity, we will only refer to the current.

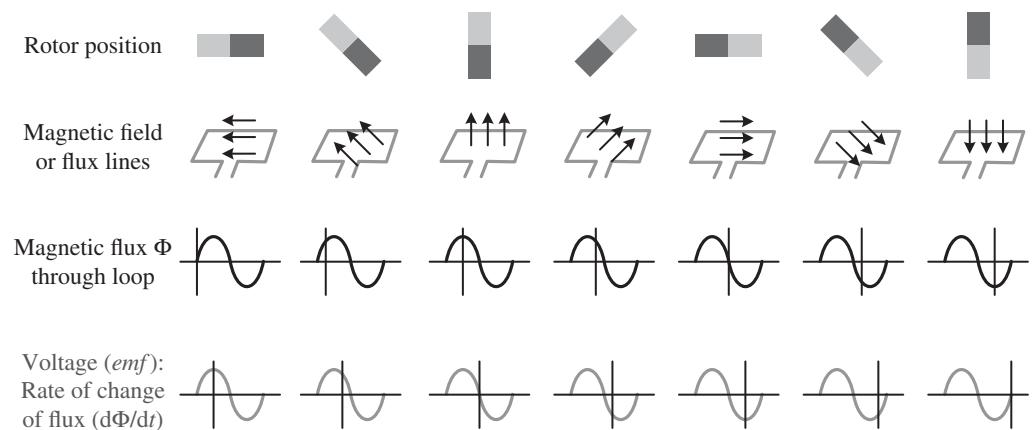


Figure 10.2 Changing flux and *emf* versus time.

remind us that because of energy conservation, any induced current will have a direction so as to oppose, not enhance, the changing magnetic field that created it. Regardless of direction, the *emf* is greatest while the flux is *changing* the most, which actually occurs at the moment when the value of the flux is zero. As the flux reaches its maximum (positive or negative), its value momentarily does not change (in other words, the very top of the sine curve is flat), and the *emf* at that instant is zero. Like the magnetic flux, the induced voltage and current will also change direction during this cycle: for one part of the cycle, with the positive flux decreasing and the negative flux increasing, the *emf* will be negative (with respect to the conventional reference direction), and for the other part, with the negative flux decreasing and the positive flux increasing, it will be positive. Mathematically, this *emf* and its resulting current could be plotted as another (more or less) sinusoidal wave, but offset from the first.³ We have generated an alternating current! The frequency of this a.c. is the same as the rotational frequency of the spinning magnet; that is, one complete oscillation of voltage or current corresponds to one complete revolution of the magnet.

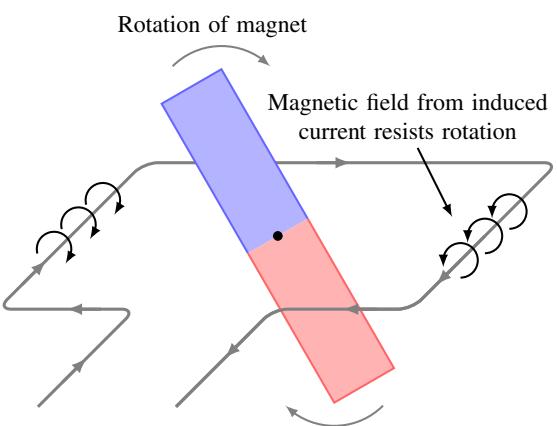
Since the wire or armature is carrying a voltage and current, it is in fact carrying electric power: a load could be connected to it, and the moving charges in the wire would do physical work while traversing that load, driven by the induction phenomenon in the generator. Thus, in abstract terms, energy is being transferred from the generator to the load. Where did this energy come from, and how did it get into the wire? From common sense, we know that energy was “put into” the generator by an external source of mechanical energy: something spinning the magnet. For a real generator, this energy source is the *prime mover*, which might be a steam turbine with a shaft connected to the generator’s “magnet.” But spinning a magnet in itself does not imply expending any energy. Rather, the magnet must offer some resistance to being spun; there must be something to push against.⁴ Where does this resistance come from?

The answer is that the magnetic field of the rotating magnet is pushing against a second magnetic field that is the result of the induced current in the wire. This is illustrated in Figure 10.3. As described in Section 1.5.3, a current-carrying wire is surrounded by a magnetic field proportional

³ Readers familiar with calculus will recall that the derivative of the sine is a cosine, or a curve of the same shape, only shifted by 90°.

⁴ In terms of physics, the mechanical work done equals the force applied multiplied by the distance an object is being pushed. If the magnet offered no resistance to being pushed, it would be impossible to exert a consistent force on it, and therefore impossible for it to do any work.

Figure 10.3 Simple generator with armature reaction.



to this current. For a straight wire, the field lines describe a circular pattern around it; when the wire is bent into a loop, these lines add together to form a magnetic field pointing straight down the middle, perpendicular to the plane of the loop. But, as we observed, the current in our wire is alternating. This results in an alternating magnetic field, changing intensity and direction in direct proportion to the current flowing; in a real generator, this field is called the *armature reaction*.

The two magnetic fields—the spinning magnet's own field and the armature reaction created by the induced current—interact with each other. This interaction is like the repulsive force that we feel when we try to push the north poles of two magnets together. It is not immediately obvious how to visualize this force changing over time as the bar magnet rotates and the induced field waxes and wanes, but we can imagine two magnets that resist being superimposed with their poles in the same direction. Thus, as the bar magnet rotates toward a field in the same direction, it will be repelled or slowed down in its rotation.

The force in this case will actually be a pulsating one, reaching a maximum at the instant of maximum current as well as a momentary zero when the current is zero. In any case, as we would predict based on considerations of energy conservation (namely, that energy cannot be created or destroyed), the force on the magnet will always act in such a direction as to retard, not accelerate, its motion.

To summarize: an external force spins the bar magnet, resulting in a magnetic field or flux that changes over time. This changing magnetic field induces an alternating voltage and current in a loop of wire surrounding the space. The a.c. in turn produces its own magnetic field, which acts to retard the motion of the spinning magnet. In this way, the interaction of the magnetic fields mediates the transfer of energy from mechanical movement into electricity. When the generator is operated in reverse as a motor, the armature field is produced by an externally supplied current, but the interaction between the two magnetic fields is completely analogous. Now, the force between the two fields acts on the magnet in the center and forces it to spin, converting electrical into mechanical energy.

10.2 D.C. Machine

In the simplest case, an electric generator produces a.c. If direct current (d.c.) is desired as the output, there is a relatively straightforward way of *rectifying* or converting the a.c. output into d.c. In this case, it is preferable to reverse the arrangement and hold the magnet stationary while

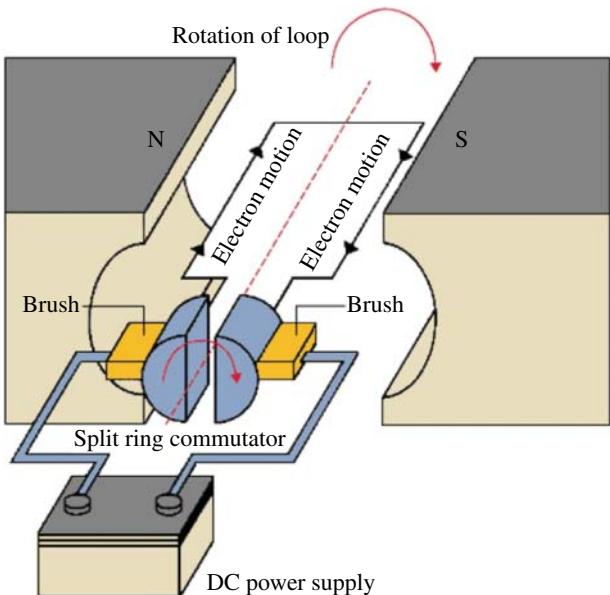


Figure 10.4 Schematic of a direct-current motor. Source: Unknown.

the armature rotates. This allows the use of a larger magnet, as illustrated in Figure 10.4. The rotating ends of the wire in which the current is generated are then connected to the load by sliding contacts, in the form of brushes or slip rings, that reverse the connection with each half turn. Thus, while the magnitude of the current still oscillates, its directionality in the terminals going out to the load always remains the same. An identical setup with sliding contacts is used to operate d.c. motors.

Historically, d.c. motors and generators were widely used before a.c. They lent themselves especially well for electric streetcars, which became a common sight in major European and Asian cities in the early 1900s (and whose abandonment in favor of automobile with internal combustion engines seems quite tragic today). One important advantage of d.c. motors is that they provide a high starting torque, which is necessary when accelerating a heavy conveyance from rest.

The key disadvantage of d.c. machines is the need for commutation, or change in polarity with every half revolution. Historically, all commutators required some form of sliding contacts, which inevitably suffer from mechanical friction, and well as from the repeated microscopic arcs associated with making and breaking of electrical contacts. More recently, solid-state technology enabled brushless d.c. motors where rapid on/off switching in semiconductors replaces the mechanical commutator.

10.2.1 The Paper Clip Motor

Building something with our own hands often gives us a new quality of insight, not to mention fun. With a few inexpensive materials, you can build your own d.c. electric motor. Feeling the (modest) torque and fiddling with your motor to get it to work well illustrates the principles of physics as no

textbook description can, and watching it actually spin is quite satisfying. Putting your paper clip motor together takes only a few minutes, and it's worth it!

Materials Needed

- 2 paper clips
- 1 small, strong magnet
- 1 C or D battery
- 1 yard of 20-gauge (AWG 20) *coated* copper wire
- 2 rubber bands or tape
- 1 small piece of sandpaper

Make a tidy wire coil by wrapping it 10 times or so around the battery. Leave a few inches of wire at both ends. Tighten the ends around the coil on opposite sides with an inch or more of wire sticking out straight from the coil to form an axle on which the coil will spin (see Figure 10.5).

Attach the magnet to the side of the battery. It may stick by itself, or you may want to secure it with a rubber band or tape. The magnet's north or south pole should point directly away from the battery (this is the way the magnet naturally wants to go).

Bend the two paper clips and attach them with a rubber band or tape to both ends of the battery so as to form bearings on which the axle rests. The clips need to be shaped so that they make good electrical contact with the battery terminals, allow enough room for the coil to spin in front of the magnet, and keep the axle in place with a minimum of friction.

After checking the fit of the axle on the bearings, use the sandpaper to remove the red insulation coating on *one* end of the wire so that it can make electrical contact with the paper clip. At the other end, remove the coating on *only half* the wire, by laying the wire coil flat down on a table and sanding only the top side. The coating left in place will interrupt the electrical contact during half the coil's rotation, which is a very crude way to approximate the effect of commutator brushes.

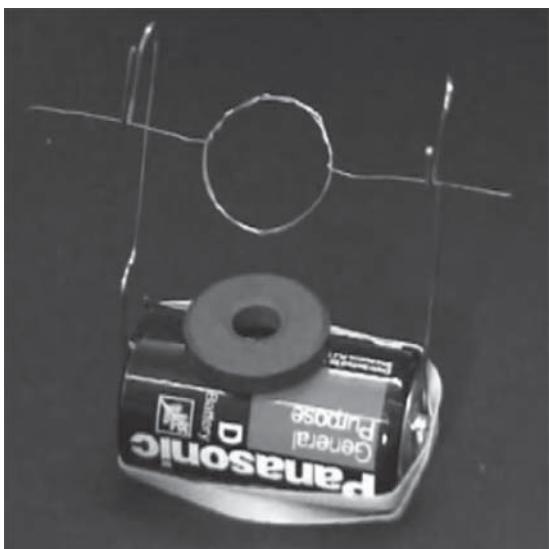


Figure 10.5 Paperclip motor.

Ideally, the direction of current flow through the coil should be reversed with every rotation, which would then deliver a steady torque on the coil in one direction; this is what commutator brushes do. If direct current were allowed to flow continuously, the direction of the torque on the coil from the changing magnetic flux would reverse with every half-turn of the coil. Simply interrupting the current for half a turn interrupts the torque during just that period when it would be pulling the wrong way. Once the spinning coil has enough momentum, it will just coast through the half-turn without power until it meets the correct torque again on the other side.

When you place the coil on the bearings with the contact side down and current flowing, you feel it being pulled in one direction by the interaction of the fields of the permanent magnet and the coil (the “armature reaction”). Now give the coil a little shove with your finger and watch it spin!

10.3 The Synchronous Generator

10.3.1 Basic Components and Functioning

Having characterized the essence of the generator’s functioning in the simplified version just discussed, we need only make some “cosmetic” changes in order to turn our description into that of a real generator. Let us begin by clarifying the nomenclature. The rotating assembly in the center of the generator is called the *rotor*, and the stationary part on the outside is called the *stator*. In the majority of designs (and the only ones considered here), the rotor contains the magnet and the stator contains the armature, which is electrically connected to the load. One simple rationale for this choice is that the armature typically carries much higher voltages, where fixed and readily insulated connections are preferable to the sliding contacts required for the rotating part of the machine. Also, making the armature the stationary outside part of the machine provides more space for those windings that carry the most current.

Our first modification of the simple generator concerns the rotor. Prior to advances in permanent magnet materials that have led to significantly stronger fields for the size and weight, permanent magnets were impractical to use for large, powerful machines. Instead, we mimic the permanent magnet by creating a magnetic field through a coil of wire (a *solenoid*, as described in Section 3.3.1) wound around an iron or steel core of high *permeability* (see Section 2.6) that enhances the magnetic field (Figure 10.6). This conducting coil is called the *rotor winding*, and its magnetic field is called the *rotor field* or *excitation field*. As will become relevant later, such an electromagnet has a big additional advantage over a permanent magnet in the rotor: namely that it is adjustable in strength. We only need to vary the amount of current flowing through the coil in order to vary the field strength proportionately. Also, electromagnets can be made in geometric configurations that create the effect of not just two, but many more magnetic poles. During basic operation, the rotor field is held constant and thus requires a constant direct current, the *rotor current* or *excitation current*, to support it. This rotor current is supplied by an external d.c. source called the *exciter* (see Section 10.3.3).

Next, consider the conductor in the stator. Instead of using just a single loop, we increase the *emf* or voltage generated in the conductor by winding it many times around successively, creating what is called the *armature* or *stator winding*. Each turn of the conductor adds another *emf* in series along the length of wire, and thus the voltages are additive. In theory, the magnitude of the generated voltage is quite arbitrary. In practice, there is a trade-off between losses (at lower voltage and higher current) and the need for insulation (at high voltage), and most utility generators operate in the neighborhood of 10–20 kV.

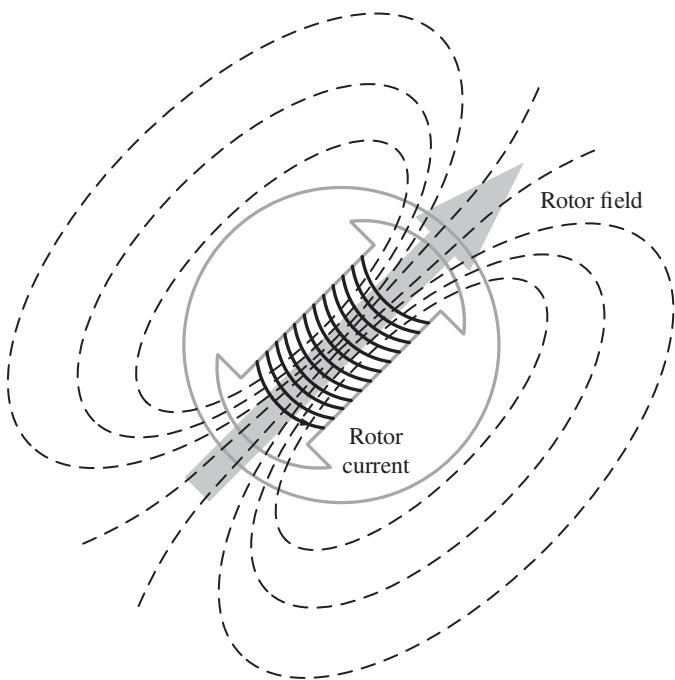


Figure 10.6 Cylindrical rotor and its magnetic field.

The many turns of the individual conductor are arranged in a staggered fashion, which makes the curve or *waveform* of the outgoing voltage and current resemble a mathematical sinusoid with reasonable accuracy. The details are beyond the scope of this text, so let us simply note that the specific arrangement of the windings through precisely machined slots in the armature is one important design aspect of generators.

Another revision to our initial model is that instead of a single conductor, a set of *three* conductors makes up the armature winding in standard equipment. These three conductors are not electrically connected to each other, but together they constitute three *phases* of a power circuit that correspond to the three wires we are accustomed to seeing on transmission lines. The phases are conventionally labeled A, B, and C. This three-phase design and its rationale are discussed in some detail in Section 4.1.

Each phase carries an alternating voltage and current offset, or shifted in time from the others by one-third of a cycle (120° , where 360° corresponds to a complete oscillation). The phases are wound such that they are also 120° apart spatially on the stator, as shown in Figure 10.7. This spatial configuration is responsible for the time delay of one-third of a cycle, since it takes the rotor that much time to pass by the given points on the armature. Aside from the engineering advantages of the three-phase system for power transmission, it provides for a much smoother conversion of power in the machine. While in our initial, single-phase version the force between the magnetic fields pulsated during the rotation, the three-phase winding provides for a uniform force or *torque* on the generator rotor. This means a better distribution of mechanical stress, and better efficiency. The same advantages apply to three-phase generators and motors.

The uniformity of the torque owes to the combination in time and space of currents in the three separate armature winding, which collectively resemble the magnetic field of a single, rotating magnet. This phenomenon can be illustrated by vector addition, or the geometric combination

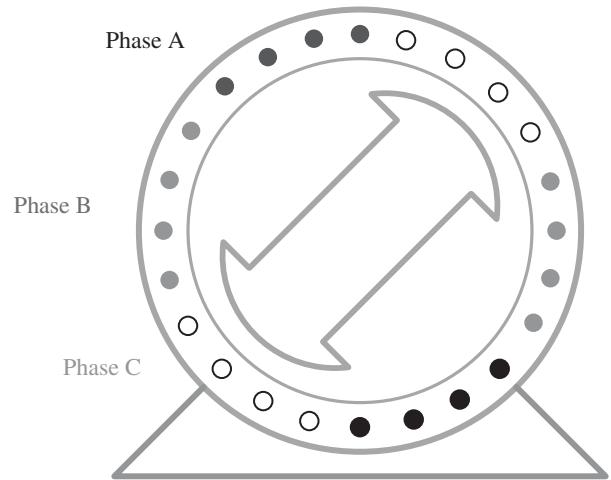
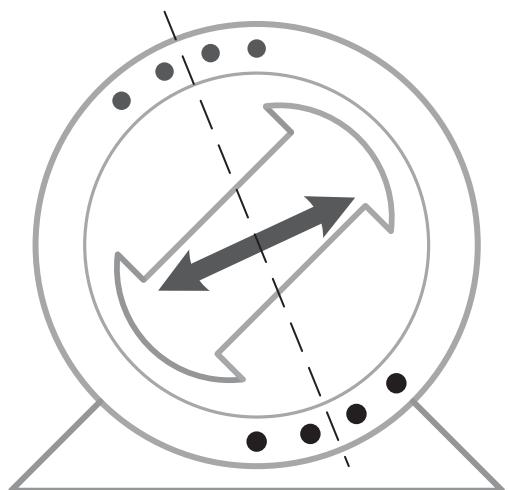


Figure 10.7 Schematic arrangement of three-phase stator winding.

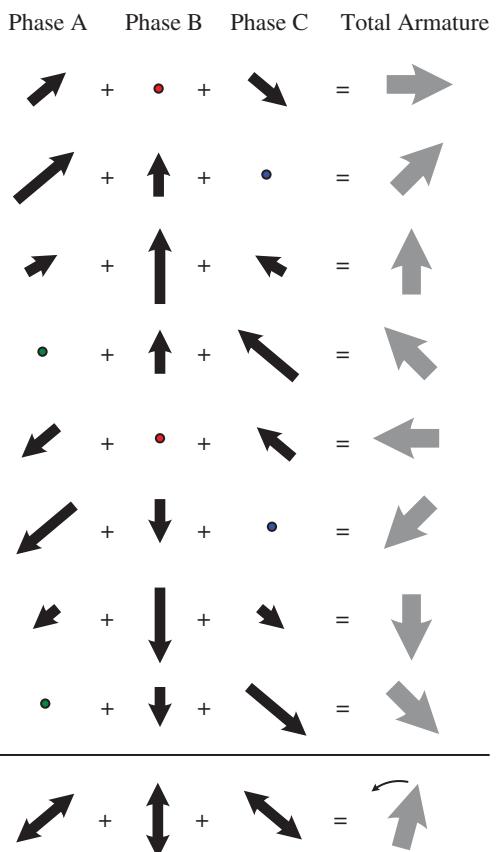


Contribution to armature reaction from Phase A

Figure 10.8 Contribution to armature reaction from one phase.

of directional quantities, where these quantities (the three magnetic fields) vary over time. Their relative timing, one-third of a cycle apart, is crucial for the result. The alternating magnetic field due to just one phase is shown in Figure 10.8. The addition of all three is illustrated in Figure 10.9, which shows the three magnetic fields of the armature windings as vectors (arrows whose length corresponds to the intensity of the field) at some distinct moments during the cycle.⁵ Owing to geometry, the sum of the three oscillating fields remains constant in magnitude, but changes direction in a cyclical fashion dictated by the relative timing of the three constituent fields: three fields stuck in place but whose magnitude oscillates in time have been converted

⁵ One cycle here is synonymous with one complete revolution of the rotor (not shown), or one complete oscillation of the alternating voltage or current, or one complete revolution of the resulting combined magnetic field that now constitutes the armature reaction.

Figure 10.9 Three-phase armature reaction.

to a single, rotating field of constant magnitude.⁶ In the case illustrated, the apparent motion is counterclockwise.⁷

The armature reaction of a three-phase generator thus appears as a steady, rotating field called the *stator field*. The stator field spins at the same frequency as the rotor, meaning that the two fields move in synchronicity and maintain a fixed position relative to each other as they spin; this is why this type of generator is called *synchronous*. As we will see in Section 10.4.2, the exact relative position of the rotor and stator fields can be adjusted operationally and relates to the generation of reactive power or the generator's power factor (Section 3.4).

10.3.2 Number of Poles

While the rate of rotation of the rotor and the frequency of alternating current in the synchronous generator are often the same, they are conventionally specified in different units: a.c. frequency is measured in hertz (Hz), equivalent to cycles per second, and rotor rotation is usually given in

⁶ For readers not familiar with vector addition, we note that the sum of two or more vectors indicates the combined effect of the fields or forces represented, simultaneously showing direction and magnitude. This sum can be obtained geometrically by placing the arrows to be added tip-to-end and connecting the origin with the final destination.

⁷ A clockwise motion would be obtained by reversing the relative timing of the three phases; for example, Phase B would reach its maximum one-third of a cycle before instead of after Phase A. This is referred to as *positive sequence* versus *negative sequence*.

revolutions per minute (rpm). Since a minute equals 60 seconds, a rotational frequency of 60 Hz corresponds to $60 \times 60 = 3600$ rpm; 50 Hz is 3000 rpm.

A complication arises in situations where such a high rotational frequency is impractical or impossible to obtain from a given prime mover or turbine. This applies especially to hydroelectric turbines, because water simply does not flow downhill that fast.⁸ The solution is to alter the design of the rotor so as to increase the rate of change of its magnetic field compared to the rotation of the entire assembly. This is accomplished by arranging the rotor windings in such a way as to produce the effect of more than two magnetic poles.

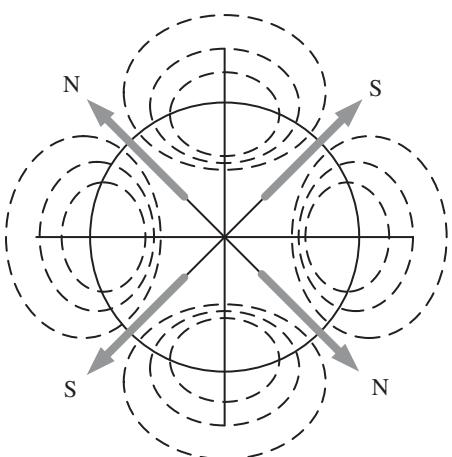
Then, each physical revolution of the rotor results in more magnetic poles moving past the armature windings, where each passing of a north and south pole corresponds to a complete “cycle” of a magnetic field oscillation. For example, if the rotor has eight magnetic poles instead of two, each armature winding is exposed to four complete cycles of magnetic field reversal during one rotor revolution. Therefore, the rotor need only spin at one-quarter the revolutions (900 rpm) in order to still produce an a.c. frequency of 60 cycles per second. This relationship is usually specified in the form of the following equation,

$$f = \frac{np}{120}$$

where f is the a.c. frequency in Hz, n is the rotational rate of the rotor in rpm, and p is the number of magnetic poles. The factor of 120 comes from 60 seconds per minute and two poles in a single magnet.

In theory, the number of poles on the rotor can be any even number, but powers of two (4, 8, 16, and 32) are especially common. Figure 10.10 illustrates a four-pole rotor and its magnetic field lines. To produce these poles, the rotor windings are configured in coils pointing outward from the center of the rotor, wound on more or less protruding cores; this type of arrangement is known as a *salient pole* rotor. Two-pole generators use a *cylindrical* or *round rotor* with rounded edges instead, whose field more closely resembles the bar magnet of our initial example (Figure 10.1). This round shape is important for minimizing drag (air resistance) at high rotational speeds.

Figure 10.10 Magnetic field of a four-pole rotor.



⁸ Another intriguing case is that of very large steam turbines where, at 3600 rpm, the tangential velocity on the outside edge of the turbine blades would exceed the speed of sound. Unlike helicopter blades, it is not practical to engineer an efficient steam turbine to withstand the associated mechanical stresses.

10.3.3 Other Design Aspects

As mentioned earlier, the terminal voltage of most large generators interconnected on the bulk grid is on the order of 10 kV. In general, a reasonable choice of generator voltage is bounded by two considerations. Higher voltage requires more separation and insulation between parts, and generally increases the hazard. Ultimately, there is also a space limitation, as the voltage is produced by many armature conductor turns in series. Low voltage, on the other hand, becomes increasingly inefficient for large amounts of power, since delivering the same power at a lower voltage requires a proportionately greater current. Current, however, is associated with resistive heating of the conductors, and thus with thermal (I^2R) losses. Aside from wasting energy, the heating of the conductors inside a generator limits its output capacity. It is therefore desirable to use the highest practical voltage level.

Even so, the conductors on both the rotor and stator still need to carry very large currents and must be designed to do so safely. For example, consider a generator producing 100 MW at 10 kV on three phases. Because $P = IV$, the current in each phase exceeds 3000 A. Since the limiting factor is the dissipation of heat, conductor diameter for lowering the resistance as well as cooling of the conductor are important. A quick calculation shows that if a 100-MW generator has an efficiency of 99%, the 1% lost represents 1 MW of heat continuously released into the generator's environment—equivalent to hundreds of residential space heaters in a single room.

While smaller generators can be passively or actively air-cooled (using fans to force the air), the cooling medium of choice for larger generators is hydrogen gas. Hydrogen conducts heat well and creates relatively little drag at high speeds. The hydrogen coolant is kept at an overpressure inside the generator so as to prevent infiltration of air, since a mixture of hydrogen and oxygen may ignite. It was the development of hydrogen cooling systems that provided for the increase in the size and efficiency of state-of-the-art generators in the 1940s and 1950s.

Like transformer windings, all the conductors inside a generator are coated with a thin layer of insulation to force the current to flow around the loops rather than short-circuiting across windings or through the rotor or stator iron. This insulating layer can be thin because, although the currents carried by these conductors are very large, the potential differences between adjacent windings are relatively small. The temperature tolerance of the insulation material sets an important limit on the currents to which the generator windings can safely be subjected, since overheating will cause degradation and eventual failure.

The geometry of generator design is primarily concerned with guiding and utilizing the magnetic fields in the most effective way possible. This involves detailed analysis of the air gap between rotor and stator, which is the actual locus of the interaction between the rotor and stator fields, where the physical force is transmitted. It becomes especially important here to consider fringe effects such as leakage flux and distortions around the edges of magnetic materials, since magnetic flux does not remain perfectly confined to the high-permeability regions. Furthermore, one must account for *eddy currents* that develop in localized areas within a conducting or magnetic material.⁹

Finally, there are several options for providing the d.c. field current to the generator rotor. For most large, synchronous generators, this is done with an auxiliary d.c. generator called the *exciter*. This exciter in turn requires its own field current, which it may draw from its own output in a

⁹ Electrical eddy currents are those that flow locally within a conductor rather than traveling its full length, much like pools of whirling water in a stream.

process called *self-excitation*,¹⁰ or it may contain a permanent magnet. While the exciter need only provide sufficient energy to overcome heating losses in the generator rotor, it is convenient to mount it on the same shaft along with the a.c. generator and the turbine and simply draw some of the mechanical turbine power for its purpose. Alternatively, a synchronous generator could draw its field current from a battery, but this is not generally practical. The third option is to draw directly upon the a.c. grid for the excitation current by *rectifying* it, that is, converting it from a.c. to d.c. Note that this latter approach only allows for the generator to be started up with the help of the external a.c. system providing the proper voltage; such a generator will not have the capability to start up in the event of a system blackout known as *black-start capability*.

10.4 Operational Control

Ultimately, the operational control of interconnected synchronous generators in a power system must be understood in terms of the interactive behavior among the generators, as discussed in Sections 10.4.3 and 10.4.4. Let us begin, though, by introducing the basic concepts for controlling an individual generator. While we distinguish many specific quantities within a generator—fields, currents, and voltages—these variables are so highly interdependent that there are only two real control variables in actual operation, or two ways to adjust a generator's behavior. These variables are the rotational frequency of the generator, which is related to the real power it supplies, and the voltage at its terminals (referred to as the generator *bus voltage*), which is related to the *reactive power* it supplies.

The term *bus* is very important in the analysis of power systems. Derived from the Latin *omnibus* (“for all”), the *busbar* is literally a bar of metal to which all the appropriate incoming and outgoing conductors are connected. To be more precise, the busbar consists of three separate bars, one for each phase. Called bus for short, it provides a reference point for measurements of voltage, current, and power flows. In analyses of networks, buses represent the nodes that must be characterized, while the detailed happenings “behind” the bus can be ignored from the system point of view. For a generator, voltage and current measurements at its bus are the definitive measure of how the generator is interacting with the grid.

10.4.1 Single Generator: Real Power

Real power output is controlled through the force or torque exerted by the prime mover, for example, the steam turbine driving the generator rotor. Intuitively, this is straightforward: if more electrical power is to be provided, then something must push harder. The rotor's rate of rotation has to be understood as an equilibrium between two opposing forces: the torque exerted by the turbine, which tends to speed up the rotor, and the torque exerted in the opposite direction by the magnetic field inside the generator, which tends to slow it down. The slowing down is directly related to the electric power being supplied by the generator to the grid. This is because the magnetic field that provides the retarding effect (the armature reaction) is directly proportional

¹⁰ It is not obvious that a generator should be able to supply its own excitation, since it cannot start to produce an excitation current without a magnetic field. The answer is that after an initial external magnetization of the “virgin” generator, a small amount of residual magnetization remains as a memory in the core material even after the generator is shut off. This small residual field is sufficient to interact in a mutually amplifying way with its own *emf* once a torque is applied, so as to reconstitute the full magnetic field.

to the current in the armature windings, while the same current also determines the amount of power transmitted.

For example, if the load on the generator suddenly increases (someone is turning on another appliance), this means a reduction in the load's impedance, resulting in an increased current in the armature windings, and the magnetic field associated with this increased current would slow down the generator. In order to maintain a constant rotational frequency of the generator, the turbine must now supply an additional torque to match. Conversely, if the load is suddenly reduced, the armature current and thus its magnetic field decreases, and the generator would speed up. To maintain equilibrium, the turbine must now push less hard so that the torques are equal and the rotational frequency stabilizes.

The torque supplied by the prime mover is adjusted by a governor valve (Figure 10.11). In the case of a steam turbine, this increases or decreases the steam flow; for a hydro turbine, it adjusts the water flow.¹¹ This main valve can be operated manually (i.e., by deliberate operator action) or, as is general practice, by an automated control system. In any situation where a generator must respond to load fluctuations, either because it is the only one in a small system or because it is designated as a *load-following* generator in a large power system, automatic governor control will be used; in this case, the generator is said to operate "on the governor."

As discussed further in Chapter 11, the automatic governor system includes some device that continually monitors the generator frequency. Any departure from the set point (e.g., 3600 rpm) is translated into a signal to the main valve to open or close by an appropriate amount. Alternatively, a generator may be operated at a fixed level of power output (i.e., a fixed amount of steam flow), which would typically correspond to its maximum load (as for a *base-load* plant); in this case, the generator is said to operate "on the load limit."

Various designs for governor systems are in use. Older ones (like the classic Watt governor in Figure 11.1) may rely on a simple mechanical feedback mechanism such as a flywheel that expands

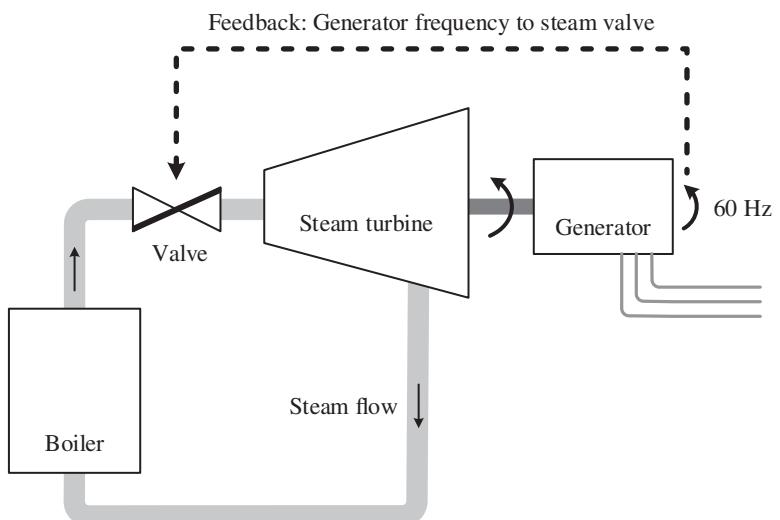


Figure 10.11 Controlling generator output with the governor valve.

¹¹ It is usually a good approximation to say that torque depends only on flow rate, which assumes a constant pressure of the steam or falling water. In reality the pressure would tend to vary inversely with flow rate, but not much.

with increasing rotational speed due to centrifugal force, which is then mechanically connected to the valve operating components. Modern governors are based on solid-state technology and digitally programmed, providing the ability to control based on not just the frequency measured in real time but its time rate of change (i.e., the slope). This allows anticipation of changes and more rapid adjustment, so that the actual generator frequency ultimately undergoes much smaller excursions. In any case, such a governor system allows the generator to follow loads within the range of the prime mover's capability and without direct need for operator intervention.

10.4.2 Single Generator: Reactive Power

The other dimension of generator control has to do with voltage and reactive power, which are controlled by the field current provided to the rotor windings by the exciter. This effect is rather less obvious than the relationship between rate of rotation and real power, and to understand it we must further discuss the magnetic fields within the generator.

To begin, we can readily appreciate that increasing the d.c. current in the rotor will result in a corresponding increase in the rotor magnetic field, which in turn increases the electromotive force in the armature's windings.¹²

The important point to realize here is that a generator's *emf* first manifests as a voltage, or potential difference between the generator's terminals, and that this *emf* is what causes current to flow through the windings and to the load. But the magnitude of this current is determined by the load impedance. Therefore, we must think of the armature current as reflecting what happens in the load, not the generator itself. The terminal voltage, on the other hand, is driven by the generator and is almost completely independent of the load: although a very large load might cause a noticeable voltage reduction due to Ohm's law, an absence of load does not act as a voltage source. Thus, the generator terminal voltage is a function of the *emf* in the armature, which in turn is controlled by the d.c. rotor field current. Increasing the field current will increase the generator bus voltage, and decreasing the field current will decrease generator bus voltage.

What does this have to do with reactive power? Here we must consider in more detail the geometry of the magnetic field and the relative timing or phase difference of the alternating voltage and current.¹³ Let us construct a graph that shows the variation of the stator voltage and current over time in relation to the rotation of the rotor, as in Figures 10.12 and 10.13.¹⁴ We can represent these phenomena for the entire stator by referring only to a single phase, even just a single winding, in the armature. The horizontal axis indicates time, marked in terms of the rotor angle (i.e., the position of the rotor at any given instant). The interval depicted, ranging from 0° to 360°, thus corresponds to one complete revolution of the rotor. To represent other windings, we would only need to shift the zero mark on the time axis. The vertical axis indicates voltage or current, whose scales in the illustration are arbitrary since they are in different units.

The first curve shows the magnetic flux (the rotor field) as seen by the armature winding. This flux varies sinusoidally, with its maximum occurring at the time that the rotor field is pointing in a direction perpendicular to the plane of the winding. The second curve shows the (negative) rate

¹² Of course, the *emf* is a function of the *rate of change* of the magnetic flux, which will increase with greater magnitude of the rotor field. This rate of change could also be increased by a higher rotational frequency (as it is in the induction generator; see Section 10.6.1), but for the synchronous generator we assume frequency to remain constant during normal operation.

¹³ As the reader will recall from Section 3.4.2, reactive power refers to the transfer of power back and forth between circuit components that is associated with a shift in the relative timing of voltage and current.

¹⁴ Here we show quantities explicitly in the time domain, which is more intuitive for many readers. Section 10.7 describes the same phenomena more concisely in the standard format, with phasors in the complex plane.

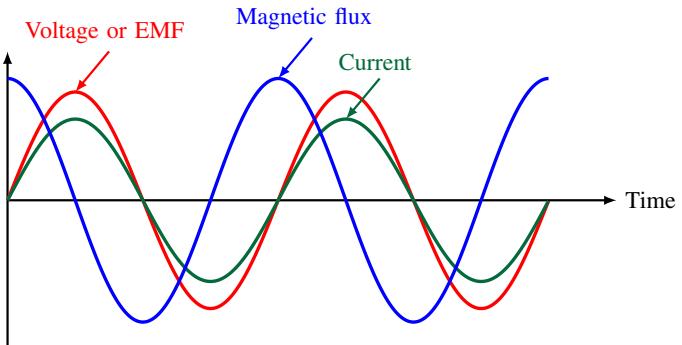


Figure 10.12 Flux, armature voltage, and current versus time; unity power factor.

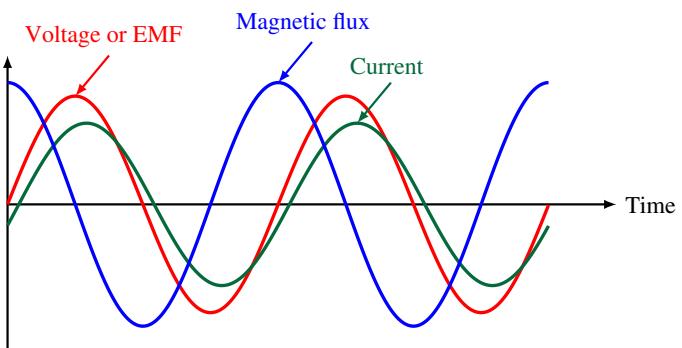


Figure 10.13 Flux, armature voltage, and current versus time; lagging power factor.

of change of this flux, which can also be read as the *emf* or voltage in the armature winding. The third curve shows the current in the armature winding, which also can be read as the magnitude of the (stator) magnetic field due to this one winding. The armature current, including its temporal relationship to the voltage, is determined by the load connected to the generator. In Figure 10.12, the load is purely resistive, or the power factor unity. Consequently, the timing of current and voltage coincides. Figure 10.13 shows a more typical operating condition with a lagging power factor, where the load includes some inductive reactance (such as electric motors). The lagging power factor means that the current lags or is delayed with respect to the voltage. Accordingly, the third curve is shifted here from the second curve, so that its maximum occurs some fraction of a cycle later. In this situation, the generator is said to “supply” VARs or reactive power to the load.

The opposite situation, shown in Figure 10.14, is less often encountered in practice. It corresponds to a leading power factor, where the generator is “taking in” VARs. Here the load appears capacitive rather than inductive, and consequently, the current peaks somewhat *ahead* of the voltage.

Let us now graph the same phenomena in a different way, one that accounts for the whole armature and takes advantage of the synchronicity, or the fact that events in the armature remain in step with the rotor’s rotation. Rather than charting voltage and current over time, as observed at a particular location (one winding), we chart the position of the rotor and stator magnetic fields in space, as observed at one instant in time. This perspective allows the use of *direct* and

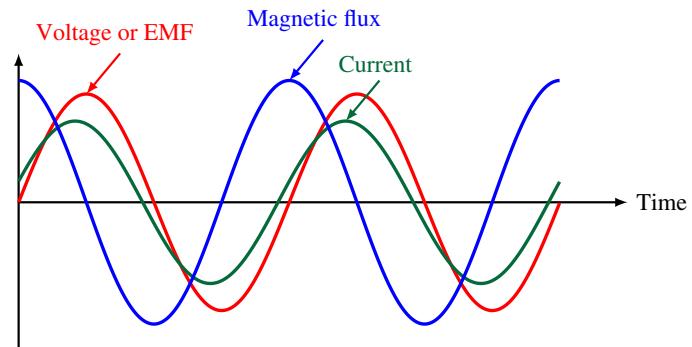


Figure 10.14 Flux, armature voltage, and current versus time; leading power factor.

quadrature components (see Section 4.3), which express time-varying quantities relative to the rotating reference frame of the rotor position.

Recall that the alternating magnetic fields of the stator (armature) windings combine geometrically so as to resemble a single, rotating magnetic field of constant magnitude. From the point of view of any given winding, the maximum effect of the rotor magnetic field occurs at a specific time interval after the rotor field has “passed by” because that is when its maximum rate of change is observed. At unity power factor, this time interval corresponds to 90° , or one-quarter of a cycle. With a lagging power factor, the time interval will be somewhat longer than 90° , since the armature current experiences an additional delay due to the reactive load; with a leading power factor, it will be somewhat shorter than 90° , since the armature current is accelerated with respect to the induced *emf* by the capacitive load. However, no matter what the angle between rotor and stator field, this angle remains fixed as both fields rotate in synchronicity. Thus, we can adequately characterize the situation by drawing two arrows representing the two fields at any arbitrary moment during their rotation. This is done in Figures 10.15–10.17 for unity, lagging, and leading power factors, respectively.

The significance of the stator field’s angle in relation to the rotor field is that it affects the amount of physical force or torque exerted on the rotor. This force is greater the more perpendicular the fields are to each other; if they were pointing in the same direction, there would be no force between them at all. It therefore makes sense, for analytic purposes, to decompose the stator field into two components, that is, to treat the stator field as the sum of two separate phenomena that have distinct

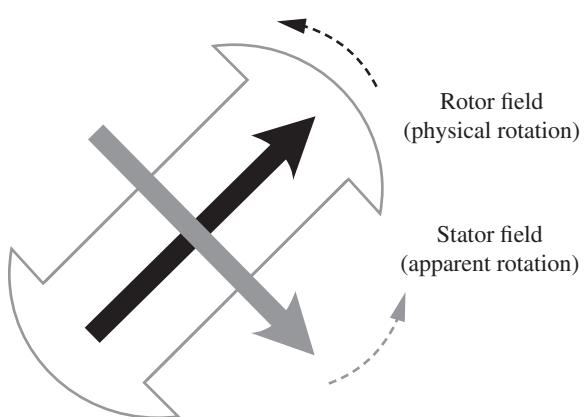


Figure 10.15 Rotor and stator field, unity power factor.

Figure 10.16 Rotor and stator field, lagging power factor.

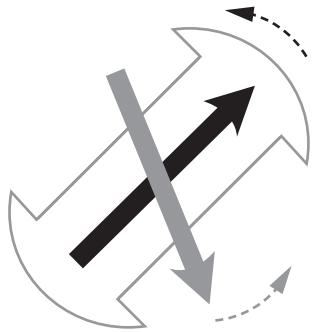
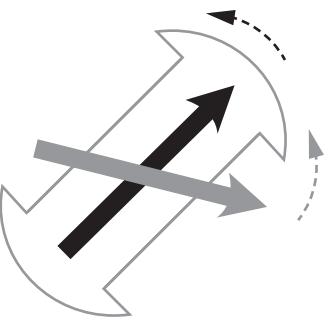


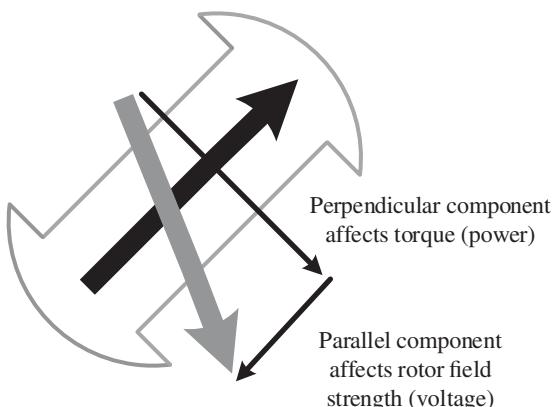
Figure 10.17 Rotor and stator field, leading power factor.



physical effects: one component that is parallel to the rotor field, and one that is perpendicular (“direct” and “in quadrature”). This decomposition is easy to perform graphically. We need only draw two arrows, one perpendicular and one parallel to the rotor field, which, when placed tip-to-end, add up to the original stator field. The process is illustrated in Figure 10.18 for the case of lagging power factor.

The perpendicular component represents the extent to which the rotor field is physically acting to exert force on the stator field, thereby enabling the conversion of mechanical power. The parallel component, on the other hand, is the one that acts in conjunction with the rotor field. Thus, the parallel component of the rotor field effectively represents an *addition* to the stator field. In the case of a lagging power factor, this is a negative addition (subtraction) since the parallel component of the stator field points in the opposite direction. Thus, when the power factor lags and the generator

Figure 10.18 Decomposition of the stator field.



is supplying reactive power, the stator field acts in part to diminish or weaken the rotor field. For a leading power factor, the stator field acts in part to increase the rotor field.

As we said earlier, the magnitude of the rotor field determines the induced *emf* in the armature, and thus the generator bus voltage. This establishes the connection between reactive power and voltage. Suppose a single generator is supplying a resistive load, and suddenly some purely reactive, inductive load (an ideal inductor) is added. There is no change in the real power (watt) requirements, only an increase in reactive power (VAR) drawn. The armature current (determined by the load) now has added to it a component that lags 90° behind its voltage and previous current. This addition amounts to a slight increase in the overall magnitude as well as an overall phase shift of the armature current in the lagging direction. Consequently, the stator field takes a new position with respect to the rotor field, now lagging behind it by some angle greater than 90°. Since the stator field grows slightly in magnitude (owing to the increased current), its perpendicular component (related to real power) does not change, but now it also has a component parallel and opposite to the rotor field, diminishing it. This diminished rotor field now produces a lower generator voltage.

An analogous process occurs if we add a capacitive instead of an inductive load. If, before the addition, the load was purely resistive, then the chain of events is as follows: the armature current has added to it a current 90° ahead of its voltage and previous current, which constitutes an overall increase in magnitude and shift in the leading direction; the stator field is now less than 90° ahead of the rotor field and thus acts to strengthen the rotor field; the generator voltage increases; the power factor is now negative (leading), and the generator is said to consume reactive power (supplied by the load). It should be pointed out, though, that this hypothetical situation is rarely encountered in practice, since most actual loads are inductive in nature. Thus, the addition of a purely capacitive load in a real situation would probably just compensate for larger, inductive loads in the system and therefore act to move the power factor closer to unity: armature current would be reduced, the stator field more perpendicular to the rotor field, and generator voltage increased.

Without any further action, a change in generator voltage will result in a change of real power supplied to the load. For example, if inductive load is added, the generator voltage drops, and with that real power would also decrease. But here, finally, is where operator action comes in. In response to the lower bus voltage, the operator or an automatic control system will increase the rotor field current (the exciter current) to compensate for its diminution by the stator field and return the voltage to its original value. Conversely, if generator bus voltage increases due to a reduction in reactive load, the field current is appropriately reduced by operational control.

Most large generating units are operated with an *automatic voltage regulator* (AVR) that maintains the generator bus voltage constant at a particular set point and continually varies the field current, as required by the load. Many large modern generators also feature an additional control loop called a *power system stabilizer* (PSS), designed to dampen any voltage oscillations that might occur (Section 13.4); this is well beyond our scope. Our focus here concerns the essential relationship between bus voltage magnitude and reactive power.

When it becomes desirable to make a deliberate change in the field current (e.g., to reallocate various generators' reactive power contributions), the operator typically intervenes by changing the voltage set point and allowing the automatic mechanism to adjust the field current accordingly. It is important to recognize that in either case, reactive power and generator voltage cannot be controlled independently. Given a particular voltage that we choose for a generator to supply, the VARs produced by this generator (and thus the generator power factor) are dictated by the load.

To a first approximation, the rotor field current consumes no energy, and changing it therefore has no direct bearing on the amount of real power generated. In reality, of course, there are thermal losses associated with rotor and stator currents as well as the current on transmission lines, and

this lost energy has to come from somewhere. In this way, an increased reactive power requirement implies a small increase in real power generation, though not necessarily by the same generator supplying the reactive power. It is important to distinguish these second-order effects from the essential relationships between real power and frequency on the one hand, and reactive power and voltage on the other hand.

A more important practical complication arises if a generator is operating at its thermal limit, and its real power output would need to be curtailed in order to safely accommodate the additional armature current associated with reactive power (see Section 10.5 on operating limits). Reactive power therefore does ultimately bear on real power generation, with some associated monetary cost.

10.4.3 Multiple Generators: Real Power

In Sections 10.4.1 and 10.4.2, we considered an individual generator supplying a load and discussed the operational changes necessary to accommodate varying real and reactive power demands. In large power systems, however, the interconnection of many generators substantively affects their operation. Any change made to any individual generator, whether torque or rotor current, has repercussions for all other generators operated synchronous with the grid. We will first examine these generators' relationship in terms of real power, which has to do with rotational frequency and their synchronicity, and then, in Section 10.4.4, turn to the problem of how reactive power is allocated among generators, which relates to their respective bus voltages.

All interconnected synchronous generators in an a.c. system not only rotate at the same (electrical) frequency, but are also in step with each other, meaning that the timing of the alternating voltage produced by each generator coincides very closely. This is a physical necessity if all generators are simultaneously to supply power to the system. As we will show, if any one generator speeds up to pull ahead of the others, this generator immediately is forced to produce additional power while relieving the load of the others. This additional power contribution results in a stronger armature reaction and greater restraining torque on the turbine, which tends to slow down the generator until an equilibrium is reached. Conversely, if one generator slows down to fall behind the others, this change will physically reduce this generator's load while increasing that of the others, relieving the torque on its turbine and allowing it to speed up until equilibrium is again reestablished.

Equilibrium here means that a generator's rotational frequency is constant over time, in contrast to the transient period during which the generator gains or loses speed. We assume that all generators will quickly settle into such an equilibrium following a disturbance. But while still at the same frequency, the new steady state position may be shifted ahead or behind in terms of phase (i.e., the exact instant at which the maximum of the generated voltage occurs). This steady-state difference between the relative timing of voltages as supplied by different generators (or measured at different locations in the grid) is referred to as the *power angle* (Figure 10.19) or *voltage angle*, often denoted by the symbol δ (lowercase delta). Each generator's power angle is directly related to its share of real power supplied: the more ahead the power angle (expressed as a greater positive angle), the more power the generator is producing compared to the others.¹⁵

The power angles can only vary by a small fraction of a cycle, or else synchronicity among the generators is lost. This problem is referred to as *stability* (see Section 13.4). For interconnected generators, loss of synchronicity means that the forces resulting from their electrical interaction

¹⁵ This relationship is consistent with the observation from power flow analysis (Chapter 12) that real power flows from greater to lesser voltage angle, assuming a predominantly inductive transmission network.

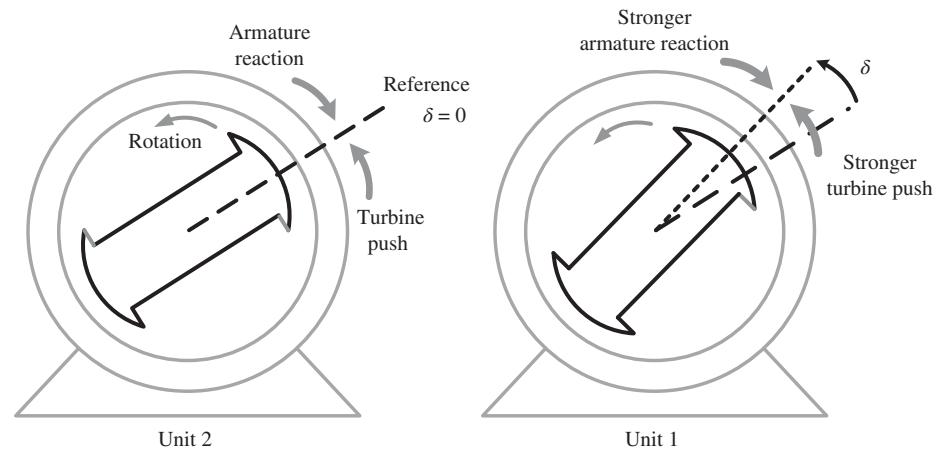


Figure 10.19 The power angle.

no longer act to return them to a stable equilibrium, as we assume in the present discussion, making their coordinated operation impossible. In terms of managing the grid, such a condition spells disaster; fortunately, for most systems under normal circumstances, synchronicity naturally follows from the inherent physics of generators.

We now examine why an increase in one generator's phase angle corresponds to an increase in real power delivered to the system by that generator, and what the forces are that tend to keep the generators in sync. For simplicity, we shall only consider the interaction with one nearby generator, not the system as a whole. This approximation is especially good for two generating units at the same power plant. It is justified physically in that the impedance between two very nearby generating units is small compared to the impedance of the remaining system and the load (i.e., the path between them is a preferred path for any current), meaning that these units will interact more strongly with each other than with the rest of the system. Although more complicated, the same analysis can in principle be applied to the entire system. We refer here to a hypothetical case of two generators, Units 1 and 2, at Plant A and observe their interaction as the power generation level on Unit 1 is increased.¹⁶

We begin by increasing steam flow at Unit 1. The additional forward torque on the turbine causes it to speed up. This acceleration of the rotor causes it to move slightly ahead compared to the rotor in Unit 2, meaning that the maxima of the *emf* or voltage produced in each phase of the armature windings at Unit 1 occur slightly ahead of those in Unit 2: the voltage or power angle δ is increased slightly. Note that the two units are still considered *synchronous* in that they remain in step with each other and their movement remains interdependent; only one marches a small fraction of a step in front of the other. Indeed, the power angles would be identical only if both units supplied the same amount of power to a load exactly in the middle.

This change in timing of the voltage results in a net difference between the voltage of Unit 1 and the system, represented by Unit 2, as measured at any given instant. Let us refer to this difference between the slightly ahead (V_1) and normal (V_2) voltages as the “difference voltage.”¹⁷

¹⁶ The terms *unit* and *generator* are synonymous here, since a generating unit refers to a single turbine-generator assembly.

¹⁷ “Difference voltage” is not a technical term, but helpful for building intuition.

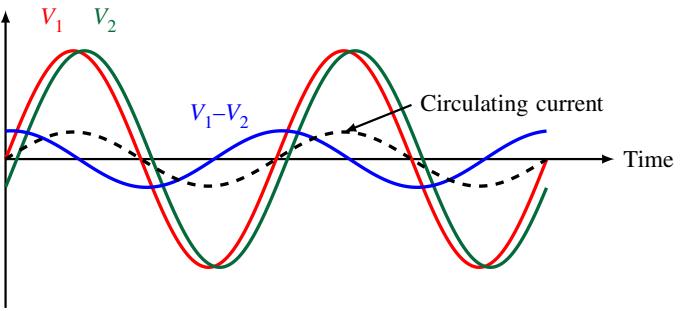


Figure 10.20 “Difference voltage” due to phase angle shift, resulting in a circulating current in phase with voltage that affects real power exchange.

If we graph this difference voltage over time, we obtain a curve that is approximately 90° out of phase with both (see Figure 10.20).¹⁸ From the perspective of Unit 1, this difference voltage is positive at those times when V_1 (its own voltage) is greater than V_2 , and negative when V_1 is less than V_2 ; this is shown in the figure. From the perspective of Unit 2, of course, we would draw the difference voltage just the opposite way and call it negative when V_1 is greater than V_2 , positive when V_1 is less than V_2 .

The difference voltage is associated with a current that flows between the two generators. This current is called a *circulating current* because it essentially circulates between the two units. (There is also a circulating current between Unit 1 and the other generators in the system, but we assume it is smaller because of the higher impedance between them; this is why representing the whole system only by Unit 2 is a reasonable approximation.) The timing of this circulating current is crucial. We must recognize two things: First, because Units 1 and 2 have opposite perspectives (they see a voltage of opposite sign), the circulating current measured in the armature windings of each generator will be opposite. Casually speaking, we could say that the current is leaving one side as it enters the other. Second, we recognize that the impedance of the circuit comprising two nearby generators is essentially all inductive reactance and very little resistance, since the wires are thick and coiled. Therefore, the circulating current will lag by almost 90° the difference voltage that induced it to flow.

The result, as shown in Figure 10.20, is a circulating current that is approximately in phase with the regular initial current generated by Unit 1, but approximately 180° *out of phase* (just opposite) with the same initial current generated at Unit 2. Thus, the circulating current *adds* to the armature current at Unit 1, but *subtracts* from the same current at Unit 2. Accordingly, one would observe an increase in the magnitude of the armature current at Unit 1 and a decrease in armature current at Unit 2. Since the magnitude of the generator voltage has not been changed in either case, and since power is the product of voltage and current, this means that the (real) power output of Unit 1 increases, while the power output of Unit 2 decreases. Effectively, Unit 1 pulling ahead in phase means that it will “push harder” and literally take some of the load off of Unit 2. In fact, the load reduction will be shared among all the generators in the system, with the most significant change in the closest units.

It is important, then, to recognize that all interconnected synchronous generators will be affected in principle by an increase or decrease in power output of any one generator. Suppose that, in our

¹⁸ The smaller the power angle or relative phase shift, the closer to 90° will the difference curve be out of phase with the two originals. In practice, the power angle should be a small fraction of a cycle; this relates to angle stability discussed in Section 13.4.

example, Unit 1 pulls ahead of the others by increasing its turbine power output, while no other changes are made in the system. As we have shown, circulating current between generators will flow so as to take load off the other generators. In response to the reduced mechanical resistance, these generators will tend to speed up. It is reasonable to assume that their governor systems sense the increasing speed and accordingly reduce the turbine power output. However, if no such corrections were made, the frequency of the entire system would increase as a result of total generation exceeding the total load. Conversely, if the total power generated were less than the load, all the generators would slow down.¹⁹

When a synchronous generator is connected to an energized system that is already operating at the specified frequency, a process of *synchronization* is required, also referred to as *paralleling* the generator to the grid. In this process, the generator is first brought up to its synchronous speed while still electrically disconnected. With instrumentation on both circuits, the frequency as well as the relative phase of the generator and the rest of the system are carefully compared, and small adjustments made on the generator speed to match the phase precisely. Once the match is achieved, the electrical connection is established by closing a circuit breaker between the generator and its bus.²⁰ Finally, with the generator “floating” at zero load, turbine steam flow and field current are increased until the generator is delivering its specified power output.

10.4.4 Multiple Generators: Reactive Power

From the principle of energy conservation, it is clear that the total amount of real power supplied by a set of interconnected generators is dictated by the load: since energy is neither created, destroyed, nor stored (in appreciable quantity) within the transmission system, the instantaneous supply must equal the instantaneous demand. If necessary, this principle will enforce itself: if operators tried to generate a different amount of power than is being consumed, the system’s operating state—first frequency, and ultimately voltage, too—would change so as to make energy conservation hold true.

The energy associated with reactive power is similarly conserved. Although reactive power involves no *net* transfer of energy over time from generators to load, the instantaneous flow must still be accounted for. Specifically, the energy going into the electric or magnetic field of one device in a circuit during some part of the cycle must be coming out of some other device, which stores that energy during the complementary part of the cycle. Therefore, while the nomenclature of “generating” and “consuming” reactive power is an arbitrary convention, it remains a useful operational reference for ensuring instantaneous power balance. Like real power, the total reactive power or VAR output of a set of interconnected generators is dictated by the load. The conservation of reactive power will enforce itself through changes in voltage.

Just as the allocation of real power generation among interconnected generators can be varied by means of the power angle or relative timing of the voltage, so can the allocation of reactive power. Here, the means of control is the magnitude of the supplied voltage in relation to other generators, which is in turn controlled by the rotor field current. If all generators in a system were

¹⁹ As stated earlier, this coupling among interconnected synchronous generators can be broken (losing synchronicity) if an excursion from the equilibrium is too large. The precise value of the power angle at which synchronicity is lost depends on a combination of generator and system characteristics; this problem is discussed in Sections 7.3.2 and 13.5.

²⁰ Today, this paralleling operation can be handled with automated controls. Traditionally, it was performed by a human operator with the aid of an analog dial called a *synchroscope*, rotating at the difference between generator and system frequency. The operator would drive the generator speed to slow down the synchroscope’s rotation, and manually close the breaker just when the dial points to 12 o’clock, indicating phase alignment. A lapse in hand-eye coordination could do serious damage to the generator.

generating the same bus voltage, their power factors would be approximately the same,²¹ and thus their reactive power output would be in the same proportion to the real power they generate. It may be desirable to change this balance for any of a number of reasons: to maintain a certain voltage profile throughout the system; to minimize cost (because one unit might generate additional VARs at lower cost than another); or because some generators are operating at their capability limit and can only produce more VARs at the expense of real power. In general, though, it is not desirable to maintain a gross imbalance between VAR generation at different units, because, as we will show, a circulating current is associated with differences in voltage levels among generators, and this current entails losses in the generator armature windings and transmission lines.

To understand the interdependence of generator voltage levels and reactive power contributions, consider again two nearby generators, Units 1 and 2. Suppose also that we wish to increase Unit 1's contribution to the system's reactive power needs, while reducing that of Unit 2. This change is effected by increasing the rotor field current to Unit 1, which in practice would be accomplished by raising the voltage set point. The situation is illustrated in Figure 10.21 with V_1 slightly greater than V_2 .

Again, we can graph the change in voltage as a difference voltage, but now this difference is perfectly in phase with the "normal" voltage generated, and the change is therefore observed simply as an increase in the magnitude of that voltage. The increased voltage results in an increased current in the armature windings of Unit 1. To a first approximation, this current circulates in the local circuit composed of the two neighboring generator stators and the bus connecting them.

As before, we note that since the impedance of this circuit is almost all inductive reactance and no resistance, the circulating current is 90° out of phase with the voltage. Thus, it has no effect on the real power generated by either generator. Also, since Unit 2 has an "opposite" perspective, the circulating current in its armature is negative at the same time as it is positive in Unit 1. In Unit 1, the circulating current is lagging and thus coincides with the lagging component of the armature current that is associated with reactive power supplied to the load. In Unit 2, on the other hand, the same circulating current is observed as a leading current (since being just opposite or 180° apart brings it from 90° lagging to 90° leading).

Recall now that the magnetic field associated with a lagging or leading armature current acts to weaken or strengthen the rotor field, respectively. Thus, the circulating current weakens the rotor field in Unit 1 and strengthens that in Unit 2. Assuming that all units were operating at a somewhat lagging power factor to begin with, the magnitude of the armature current in Unit 2 is now less, since it is relieved of some of its lagging component by the leading circulating current. As a result, the reactive load on Unit 1 is now increased, while Unit 2 experiences a decrease in reactive load.

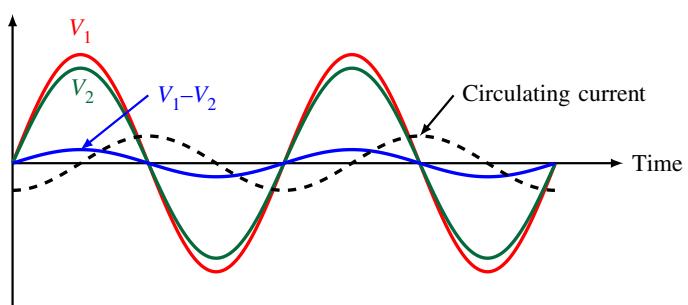


Figure 10.21 Circulating current with Unit 1 providing more reactive power.

21 Not accounting for the locational effects of the transmission network.

Let us suppose that the automatic voltage regulators in both units are out of service, and all changes to field current are made manually. In Unit 1, operators increase the field current to initiate the chain of events. The generator bus voltage increases, but not as much as one would have expected before taking the circulating current into account. Indeed, the higher they attempt to raise Unit 1's voltage compared to other units, the more field current will be required in order to achieve a further increase in rotor field strength and thus voltage. This condition is synonymous with supplying more reactive power. In Unit 2, the voltage also increases because the rotor field is strengthened by the circulating current and less reactive power is supplied. In response, the Unit 2 operators may lower the field current so that the voltage returns to its previous value.

In practice, with voltage regulators in service, Unit 1 operators would increase the voltage set point and allow the regulator to increase the field current automatically. The voltage regulator at Unit 2 would recognize the elevated voltage due to the circulating current and reduce the field current appropriately.

Overall, we might say that the circulating current has the effect of equalizing the voltages between generators, analogous to the way that a circulating current tends to equalize the rotational frequencies and real power output between generators. Again, this discussion extends qualitatively to other generators in the system, with the closest ones being affected most. The electrical interaction between generators thus results in a stabilizing force that tends to equalize voltages, analogous to the force that tends to equalize frequency.

Let us emphasize again that the total amount of reactive power supplied by all generating units is determined by the load (including the transmission and distribution network, which is generally inductive), and that this reactive load can be shared among generators in whatever way is most economical, independent of real power allocation. Typically, because of the losses associated with circulating current, an economic allocation of reactive power will have generators operating at fairly similar power factors.

A simple example of how such an allocation might look is shown in Figure 10.22, where four units, located at two power plants, are supplying an area load. (For clarity, the example ignores real and reactive line losses.) This example highlights the fact that the *system power factor*, which is determined exclusively by the load, is different from the individual *generator power factors*. Indeed, the plant operators have no way of knowing what the system power factor is, based on observations only at their own unit.

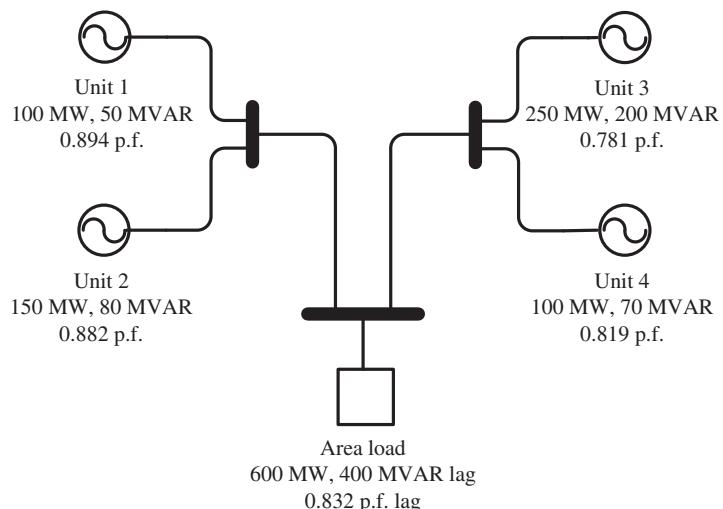


Figure 10.22 Sample allocation of reactive power among generators.

10.5 Operating Limits

The generating capacity of electric power plants is generally referred to in units of real power output, such as kilowatts or megawatts. This is appropriate in that the decisive constraint on how much power can be generated is the ability of the prime mover to deliver mechanical power to the turbine. In a typical plant design, the electric generator will be sized so as to be able to handle any amount of real power the turbine can provide.²² However, generator performance limits are still important in the operational context, because they also apply to reactive power, which is independent of the prime mover. Indeed, the appropriate measure of capacity for the electric generator as such is not in terms of real power, but rather in units of *apparent power*, kilovolt-ampere (kVA) or megavolt-ampere (MVA). It is thus a combination of real and reactive power that must be considered in order to determine whether the generator is operating within its range or is in danger of becoming overloaded.

“Overloading” for a generator primarily means overheating due to high current, though some mechanical factors may also be relevant. Excessive temperature will cause the insulating material on the generator windings to deteriorate and thus lead to an internal fault or short circuit. Different rates of thermal expansion between the winding conductors and the core at excessive temperatures can also cause insulation damage through movement and abrasion. Depending on the particular operating condition, “hot spots” may develop on different components, which is problematic because the temperature cannot readily be measured everywhere inside the generator. Possible sources of mechanical damage under excessive loading include rotor vibration due to imperfect balance, vibration due to fluctuating electromagnetic forces on the components, and loss of alignment between turbine and generator shafts due to thermal expansion or distortion of the generator frame.

Any of these types of damage are irreversible in that the generator will not recover after the load is reduced. Therefore, rather than waiting for signs of distress under high loads, generators are operated within limits specified by the manufacturer that will allow for some margin of safety to ensure the integrity of the equipment. To a first approximation, these limits are indicated by the generator *rating* in kVA or MVA. Since apparent power is directly proportional to current, regardless of the relative proportions of real and reactive power, this is synonymous with a limit on the current in the armature or stator windings.

Interestingly, there has been some historical change in the implicit conventions for such ratings, not only for electric generators. In the engineering tradition, it has long been customary to provide substantial safety margins in the design of machinery, to the extent that an experienced operator can at times exceed the nominal ratings with confidence. Yet over the past few decades, the philosophy of building some slack into technical systems by generously oversizing components has increasingly given way to a more refined approach where, aided by more sophisticated instrumentation and computing, components can be matched more precisely to needs and specifications. While such refinement in design has obvious economic justifications, it does in some sense increase the vulnerability of the system—as in the case of a generator being operated at 100% of its nameplate rating, which will now tend to have less tolerance for excursions from normal operating conditions.

²² Wind turbines are an interesting exception. Because of the dramatic variability of wind speed (along with the fact that power increases with the cube of wind speed), it is unrealistic to install a wind generator capable of handling the maximum wind power during a storm. Instead, proper design calls for an optimization that considers the expected range and frequency of wind speeds, the cost of the generator in relation to its size, and the reduction in operating efficiency that occurs when the generator is producing much less than its rated output. It is also necessary to include some type of mechanical restraint that will shut down the wind turbine beyond a certain wind speed (the *cutout speed*) and prevent it from overloading the generator.

In most situations, the current in the armature windings can be used as a criterion for the generator operating limit, as in a red line on the “generator current” display. Under certain operating conditions, however, it becomes necessary to observe more stringent criteria on generator loading, especially if there is little margin in the rating. These criteria have to do with heating of components other than the armature windings, which tends to become more prohibitive when operating at a power factor very different from unity. This comprehensive information about a particular generator is captured in a diagram called the *reactive capability curve*, which indicates a boundary on permissible combinations of real and reactive power output: all points inside the area bounded by the curve are achievable without risk of damage, and all points outside this area are prohibited.

The curve consists in part of the circle that describes a constant amount of apparent power, as in the kVA or MVA rating.²³ Within the normal operating range, where the power factor is relatively close to unity, this circle does indeed prescribe the operating limit, imposed by the resistive (I^2R) heating of the armature (stator) conductors as a function of apparent power. For the typical case shown in Figure 10.23, this operating range is defined between power factors of 0.85 lagging and

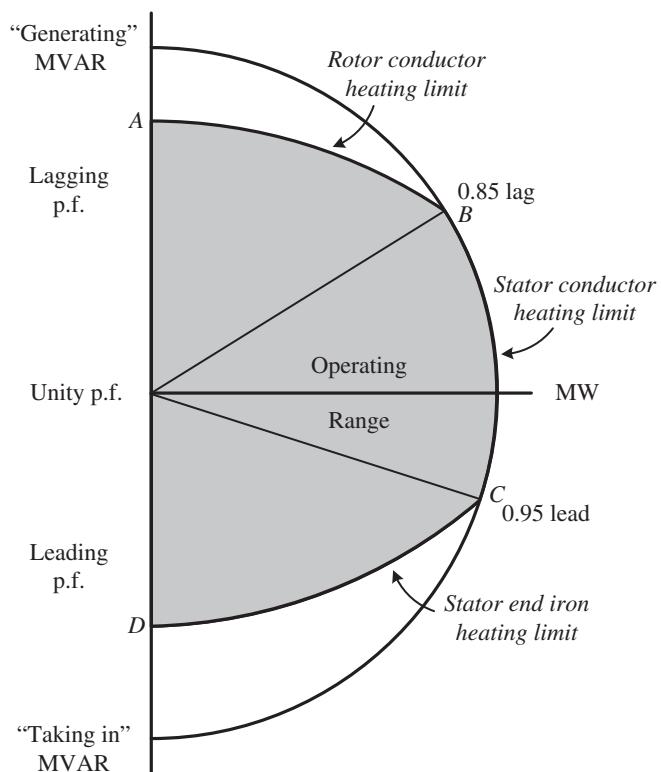


Figure 10.23 Reactive capability curve.

²³ Recall from Section 3.4.3 that apparent power can be obtained by vector addition from real and reactive power, where real power is represented by a horizontal arrow and reactive power by a vertical one, which are placed tip to end. The length of the resulting vector indicates the magnitude of apparent power, and all the possible vectors of the same length (starting at the origin) describe a circle. Here, we are interested only in the right half of this circle that corresponds to positive real power generation; the omitted left half would correspond to operation as a motor.

0.95 leading.²⁴ The applicable segment of the reactive capability curve is that between the points labeled B and C in the figure.

For more lagging power factors, it turns out that the operating limit is more stringent than what we would expect from extrapolating the circle. Instead, the limit is described by a line closer to the origin, connecting points A and B. Physically, it arises from the rotor field current or *overexcitation* required to supply large amounts of reactive power (i.e., maintaining the desired voltage despite a large reactive load). Beyond a certain power factor, heating of the rotor conductor due to this field current becomes more constraining than for the stator conductor.

In the opposite case of *underexcitation*, where the field current is reduced in response to the leading armature current beyond a certain power factor, a more stringent limit also applies. Here the concern is again about heating in the stator, but this time because of eddy currents that tend to develop within the stator core iron. The part of the reactive capability curve that applies to this condition connects points C and D in Figure 10.23.

In general, it is the operator's responsibility to assure these limits are not exceeded. Outside the normal operating range (in this example, between power factors of 0.85 lagging and 0.95 leading), automatic controls are generally used to assist the operator. For instance, voltage regulators may provide limitations that prevent the field current from being increased or decreased beyond set limits, depending on real power output.

Generator operating limits also apply to voltage. A typical tolerance range for voltage magnitude would be $\pm 5\%$. In that case, a generator nominally producing a bus voltage of 20 kV might in practice be operated between 19 and 21 kV. Staying within this range is more a matter of policy and overall system operating strategy than of urgent technical necessity from the standpoint of equipment safety.

10.6 The Induction Machine

10.6.1 General Characteristics

An induction or *asynchronous* machine is one that operates without an independent source for its rotor field current, but in which the rotor field current appears by electromagnetic *induction* from the field of the armature current. The rotor field then interacts with the stator field to transmit mechanical torque just as it does in a synchronous machine, regardless of the fact that it was the stator field that created it (the rotor field) in the first place. This may sound like pulling yourself up by your own bootstraps, but it does actually work. The catch is that some armature current must be provided externally. When motoring, this is not an issue (since the motor has to be plugged in). But when used with a prime mover to generate power, an induction generator cannot be started up without being connected to an already live a.c. system. Another practical concern is that, as we will show, induction generators can only operate at leading power factors, consuming reactive power. For both reasons, the role of induction generators in power systems is limited.

Aside from ubiquitous induction motor loads, one important application of induction machines in power systems has been for wind turbines in the 20th century. In this application, induction generators offer some advantage because they can readily absorb the erratic fluctuations of mechanical power delivered by the wind resource; they also cost less than synchronous machines.

²⁴ In the figure, this region is bounded by straight lines at angles corresponding to the specified power factors. Any point along a straight line from the origin indicates a combination of real and reactive power in the same ratio. The horizontal axis corresponds to a p.f. of unity (1.0); the vertical axis to a p.f. of zero (no real power generated at all).

However, the lack of independent controllability is a major drawback that becomes intolerable for large contributions from wind power. Modern wind machines use power electronic converters instead (Section 15.1.4).

In terms of mechanical operation, the most important characteristic of the induction machine is that the rate of rotation is not fixed, as in the case of the synchronous machine, but varies depending on the torque or power delivered. The reference point is called the *synchronous speed*, which is the speed of rotation of the armature magnetic field (corresponding to the a.c. frequency) and also the speed at which a synchronous rotor would spin. The more power is being generated, the faster the induction rotor spins in relation to the synchronous speed; the difference is called the *slip speed* and typically amounts to several percent.²⁵ When the rotor spins more slowly than the armature speed, the machine is operating as a motor. While induction machines are usually optimized and marketed for one purpose or the other, either generating or motoring, they are all in principle reversible.

Figure 10.24 shows a curve of torque versus slip speed for a generic induction machine. Zero slip corresponds to synchronous speed, and at this point, the machine delivers no power at all: neglecting friction, it spins freely in equilibrium. This is called a *no-load condition*. If a forward torque is exerted on the rotor in this equilibrium state (say, by a connected turbine), it accelerates beyond synchronous speed and generates electric power by boosting the terminal voltage. If the rotor is instead restrained (by a mechanical load), it slows down below synchronous speed and the machine is operating as a motor. Now we call the torque on the rotor negative, and it acts to push whatever is restraining it with power derived from the armature current and voltage.

The synchronous speed of a given induction machine may be equal to the a.c. frequency (3600 rpm for 60 Hz; 3000 rpm for 50 Hz) or some even fraction thereof (such as 900 or 1800 rpm),

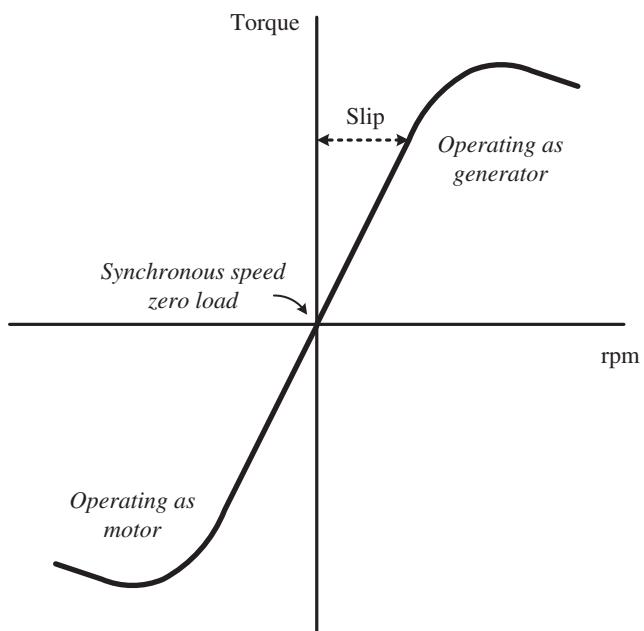


Figure 10.24 Torque versus slip for an induction machine.

²⁵ The ratio of slip speed to synchronous speed, simply called the *slip*, is also sometimes expressed as a decimal between 0 and 1.

depending on the number of magnetic poles, which in this case are created by the armature conductor windings instead of the rotor. Unlike the synchronous generator, where the stator magnetic field has two poles but the rotor field may have any even number of poles, an induction generator must have the same number of poles in the rotor and stator field because there is no independent excitation.

10.6.2 Electromagnetic Characteristics

The rotor of an induction machine consists of a set of conductors arranged in such a way that, when exposed to the armature magnetic field, a current will flow. This can be done with regular windings of insulated wire in what is called a *wound rotor*, or with a much simpler structure of conducting bars running parallel to the generator shaft that are connected in rings at the end, known as a *squirrel-cage rotor*. Squirrel-cage rotors are the most common since they are much less expensive to fabricate.

When the armature is connected to an a.c. source, it produces a rotating magnetic field just like the stator field of the synchronous generator. The rotor conductors form loops that are now intersected (*linked*) by a changing amount of magnetic flux. As the stator field rotates, the flux linking any given rotor loop is zero whenever the stator field points in a direction parallel to the plane of the loop, and reaches a maximum when the field is perpendicular to the loop. This changing flux induces a current to flow in each loop.

The timing of this current will vary from one loop to the next as the stator field makes a complete revolution. In a wound rotor, the *emfs* add together for every turn of a single conductor; rotors are typically wound for several phases. In a squirrel-cage rotor, each pair of opposite bars represents one phase. It may seem counter-intuitive that we can distinguish different currents flowing through conducting parts that are all electrically connected. In fact, they are eddy currents in a single conducting object. However, because of the object's shape, these eddy currents are very ordered, and their electromagnetic effect is similar to that of separate currents traveling along individual, insulated wires. Thus, the rotor current for each phase produces its own magnetic field, and the geometric combination of these fields into a single rotating field of constant strength works just like it did for the armature.

What is different for the rotor field, however, is the frequency. The frequency of alternating current within the rotor loops, regardless of how they are physically constructed, is given by the rate of change of the flux through the rotor loops. This rate of change depends not only on the a.c. frequency in the stator, but also the slip speed, or relative speed between the rotor's mechanical rotation and the apparent rotation of the stator field.

Suppose we connect an induction machine to an a.c. network. The rotor is initially at rest, while the stator field rotates at 60 revolutions per second (rps) (assuming a two-pole stator). This start-up condition corresponds to a *slip*, or ratio of slip speed to synchronous speed, of unity (100%). In this case, the flux linkage through the rotor loops undergoes one complete reversal with each revolution of the stator field. Accordingly, an alternating current of 60 cycles (Hz) is induced in the rotor. The resulting rotor field also rotates at 60 cycles, which implies that it remains in a fixed position with respect to the stator field, as it does in a synchronous generator.

As we will show later, the position between rotor and stator field is such that a torque is exerted on the rotor. If the rotor is initially at rest, this torque will accelerate it in the same direction of rotation as the stator field. But as the rotor accelerates toward synchronous speed, the slip decreases.

Suppose the rotor has reached 57 rps, which corresponds to a slip speed of 3 rps (or a slip of 1/20). The spinning stator field now undergoes a complete reversal with respect to the rotor only

three times per second, as the flux lines intersect the conducting loops from a different direction. Therefore, the current induced in the rotor now alternates at only three cycles.

From the perspective of the rotor, the resulting magnetic field—geometrically composed of the individual fields from each phase alternating at three cycles in different directions—appears to make a circular revolution three times per second. But how does the rotor field appear from the perspective of the stator? In this stationary reference frame, the apparent speed of the rotor field, three cycles, is added to the relative (mechanical) speed of the rotor itself, 57 cycles. The result is that the rotor field still revolves at synchronous speed, 60 cycles, just like the stator field!

The same reasoning applies for any arbitrary slip speed, positive or negative. Thus, unlike the *mechanical* rate of rotation of the rotor, which varies, the *magnetic* rotation of its field is always at synchronous speed and therefore remains in a steady relation to the rotating stator field.

As long as the rotor is mechanically spinning below synchronous speed, it experiences a changing flux in its loops, and thus has current induced in it. But as the mechanical rotor speed approaches that of the revolving stator field, the rate of change of this flux becomes less and less, as does the frequency of the induced current. Furthermore, the magnitude of the induced current diminishes, because it is proportional to the rate of change of magnetic flux. Thus, the rotor field gets weaker as the rotor approaches synchronous speed, and so does the torque between the rotor and stator magnetic fields. Finally, when the rotor reaches synchronous speed, the torque is zero.

If we now supply an external torque on the rotor, the machine speeds up beyond synchronous speed and operates as a generator. Now we have relative motion between the rotor loops and the revolving stator field, so that an alternating current is again induced. Because the relative motion is in the other direction (the rotor is revolving faster instead of slower than the stator field), the current is reversed, and the magnetic torque now acts to restrain rather than accelerate the rotor. This is the result we should expect based on energy conservation.

By forcing the rotor to maintain this faster speed, we are forcing a rotor field of a certain strength and direction to coexist with the stator field. This rotor field acts just like that of a synchronous generator: it induces an *emf* in the armature windings that will cause current to flow to the load and thus transmit electric power. Although there is already a current pre-established in the armature windings when the induction generator is first connected to the a.c. system, the induced current is additive, since the induced *emf* acts to strengthen the potential difference at the generator terminals. By contrast, when the induction machine is operating as a motor, the *emf* induced by the rotor field counteracts the existing potential difference, resulting in the motor “drawing” current from the a.c. grid.

The more the rotor speed deviates from the no-load equilibrium in either direction, the stronger the torque pushing it back toward synchronous speed. This relationship remains true up to a point, beyond which the torque diminishes (but still acts in the proper direction). This is seen in the changing slope at either end of the slip-torque characteristic in Figure 10.24. An induction generator is operated in the region between zero and maximum torque, because, as the reader may convince herself, any operating point in this region is *stable* in the sense that an excursion will be associated with a change in torque that tends to restore the operating condition. Beyond the “knee” in the curve at maximum torque, where the torque decreases with increasing speed, the operating condition is *unstable* because an increase in rotor speed further reduces the restraining torque. This is called an *overspeed* condition, which will ultimately damage the generator and must be prevented by disconnecting the prime mover if the rotor speed exceeds a given value. Correspondingly, when the machine is operated as a motor, there is an unstable condition beyond the knee at which the motor simply stops working, for example, when stopping the motor of an electric toy with your hand.

The mechanical power transmitted is given by the product of torque and angular frequency of rotation (rpm). Machines can be designed to produce maximum torque and maximum power at different operating speeds, depending on the application.

Let us now return to the question of the relative orientation of the rotor and stator fields in an induction machine, which (at least in principle) explains the constraint on the power factor. Recall that, as seen from the stator's reference frame, the rotor magnetic field is spinning at synchronous speed, and the spatial relationship between the two fields therefore remains fixed. Because there is no independent excitation current to produce the rotor field, it can only come from two sources: the stator field, and the relative movement or slip between rotor and stator. Like in Section 10.4.2, we can decompose the magnetic fields into two vector components. Again, we would decompose the rotor field into a component parallel to the stator field and one perpendicular to it. Recall also that when the rotor and stator fields are parallel, there is no torque on the rotor. This is the situation in the no-load condition with zero slip. We can then think of this parallel rotor field as the one created by the stator field.

When slip is introduced, the relative motion between the rotor and stator creates an additional component in the rotor field. This is the perpendicular component, which is associated with torque. When the machine is operated as a motor, with a mechanical force holding the rotor back, this rotor field component lags 90° behind the parallel or stator field.²⁶ When the machine is operated as a generator, with a mechanical force accelerating the rotor beyond synchronous speed, the rotor field component instead *leads* the stator field by 90° .

Because there is no way to adjust the rotor field by external means, and the only way to create the perpendicular component is through the relative motion or slip between rotor and stator field, the induction generator can only generate at leading power factor. This means that it "consumes" reactive power in the same way an inductive load does.²⁷ A more precise and unambiguous statement is that an induction machine is always *underexcited* (see Section 10.7.2). Consequently, in a power system, other generators or capacitors installed close to the induction machines must compensate for the difference and supply an appropriate amount of VARs to meet the requirements of both the induction generators and the load. By itself, an induction generator could only supply a load with capacitive reactance. This is not a relevant scenario, though, because without another a.c. source around, the induction generator could not get started in the first place.

10.6.3 Reluctance Machine

The idea of a magnetic field that advances in a rotating manner due to alternating currents in multiple windings was realized by Nikola Tesla. Figure 10.25 shows an original drawing of three-phase windings in Tesla's 1888 patent for a *reluctance motor*.²⁸ This type of machine does not have many applications in power systems, although there is a modern adaptation for electric

²⁶ This is because the rotor conductors have only inductive reactance and no appreciable resistance, and the rotor current therefore lags 90° the *emf* induced in the rotor by the stator field as it slips.

²⁷ The nomenclature can be confusing, because the terms "lagging" and "leading" are in a sense opposite for loads and generators. A generator said to operate at a lagging power factor supplies a load with lagging p.f., even though the generator is behaving like a leading load. Physically, it means that the relative timing of current and voltage are complementary in the generator and the load, so that the generator absorbs reactive power in the same instant as the load releases it. In the generator convention, current is measured in the opposite direction from load, since "positive" power is injected rather than consumed. If instantaneous current and power at the load are positive, those at the generator are negative. But flipping a sine wave over by making it negative is equivalent to shifting it by 180° , which makes "lagging" as a generator equivalent to "leading" as a load.

²⁸ This is one of many patents that are easily searchable online: <https://patents.google.com/patent/US381968A/en> (retrieved February 2024).

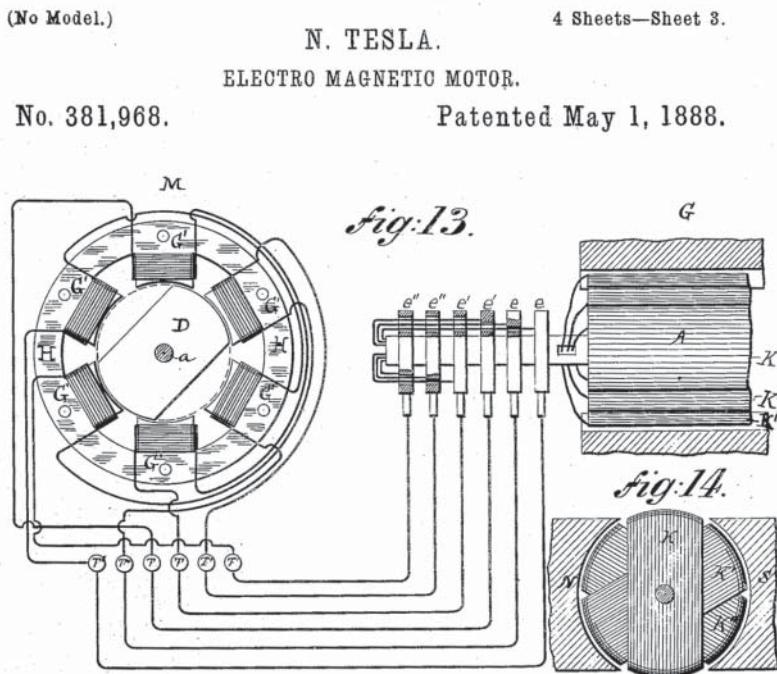


Figure 10.25 Illustration from Nikola Tesla's a.c. motor patent.

vehicle motors. It is interesting because we can think of it as a cross between a synchronous and an induction machine.

In an induction machine, the rotor field is created entirely by the armature field and moves relative to the physical rotor cage, which turns at an asynchronous speed. In a synchronous machine, the rotor field is created independently of the armature field by either a permanent magnet or an exciter. In a reluctance machine, the rotor field is initially created by the armature field, but then stays in a fixed position relative to the physical rotor, which turns at synchronous speed.

In Figure 10.25, the rotor becomes magnetized with a steady orientation in line with its cutaway shape, and experiences a torque that tends to align it with the rotating magnetic field from the three-phase alternating currents in the stator. This is analogous to a nail that gets picked up by a permanent magnet: the proximity of the permanent magnet makes the nail act as a magnet also, as the magnetic field or flux lines concentrate in the metal as opposed to the surrounding air, and the nail experiences a mechanical force as a result. In Tesla's own words, "The disk D, owing to its tendency to assume that position which embraces the greatest possible number of magnetic lines, is set in rotation, following the motion of the lines or the points of greatest attraction."²⁹

It is worth contemplating that today's cutting-edge electric motor design builds on ideas from 135 years ago, combining them with sophisticated computer-based optimization and control.

²⁹ Tesla goes on to explain that a circular disk, as indicated by the dotted lines, can also work: "This phenomenon I attribute to a certain inertia or resistance inherent in the metal to the rapid shifting of the lines of force through the same, which results in a continuous tangential pull upon the disk, causing its rotation." (U.S. Patent No. 381,968, May 1888.)

10.7 Modeling Generators

The specific behavior of electric machines under normal and disturbance conditions is its own area of study within power systems. In this section, we will introduce some basic ideas about how to analyze a machine through an equivalent electric circuit model; this will require more mathematical formalism than elsewhere in this chapter. Although the concepts apply to generators as well as motors, we mostly refer to generators.

In steady-state operation, we are interested in the exact relationships between terminal voltages and currents under different amounts of load and different power factors. It is also important to understand how a machine would respond to sudden changes in the system it is connected to, such as voltage excursions, or faults that can occur at various places in the network. For example, if there is a short-circuit in place of the normal load, how much fault current would flow through the generator windings and potentially cause damage? That type of detailed analysis is beyond the scope of this text, but we can at least get a flavor of how it would be approached.

Here we will introduce a basic electrical model for a synchronous machine, and relate this to the steady-state electrical power output of a generator. We skip a full derivation, but aim to provide a slightly more formal view of what was described qualitatively before. Specifically, we draw an electric circuit equivalent that (approximately) produces the relationships of interest between currents and voltages in the rotor and stator. This equivalent circuit model is to be understood in a manner analogous to the transformer model in Chapter 8: it does not claim to illustrate what is physically inside the device; rather, it offers a way to quantitatively link input and output variables.

We will limit ourselves to the simplest possible geometry, a synchronous machine with a round rotor. Another style of machine, the salient pole rotor, requires some modifications of the model because it takes into account the particular geometry of the rotor—specifically, how it is asymmetrical in the direction of the protruding poles. But since our objective is conceptual understanding, we'll skip the salient pole and other geometries.

We refer to the cross section of a single-phase, round rotor machine as illustrated in Figure 10.26. The crucial quantity of interest is the magnetic *flux* that links the rotor current with the stator current.³⁰ There are two contributions: that produced by the rotor, and the armature reaction. The locale where the two meet and interact is in the *air gap* between the rotor and stator. As discussed above and illustrated in Figure 10.6, the rotor magnetic field can be visualized like a simple permanent magnet, although we assume it is produced by a d.c. excitation current or field current i_F . Neglecting minor effects around the ends of the rotor windings, the flux lines are radial, perpendicular across the air gap, and rotating with the rotor.

When there is no load—that is, no current is flowing in the armature windings—the electromotive force produced by the changing flux can be measured as a voltage at the generator terminals (i.e., the ends of the armature windings). This voltage is usually labeled e or E , with a subscript like e_a for a single phase, and also referred to as the generator internal voltage. It is analogous to the voltage produced by an ideal transformer. This voltage e_a for any one winding varies sinusoidally, lagging the flux by 90° .³¹

³⁰ See Section 1.5.3 for the relationship between magnetic field and flux. When drawing a more detailed model of an electrical machine, it is necessary to think not just about the general direction in which a magnetic field is pointing (as we did earlier in this chapter, where magnetic fields appeared as straight arrows), but about the effectiveness of the interaction between magnetically coupled parts. This means we must trace the magnetic field lines as they curve around and close on themselves, to keep track of where they pass through and which components they are linking. Consequently, *flux* is a much more relevant and useful quantity here.

³¹ This can be derived using Gauss's Law and straightforward geometry for a round rotor, where the magnetic flux is integrated over a cylindrical surface. For simplicity, it is assumed that the many turns of each phase winding are

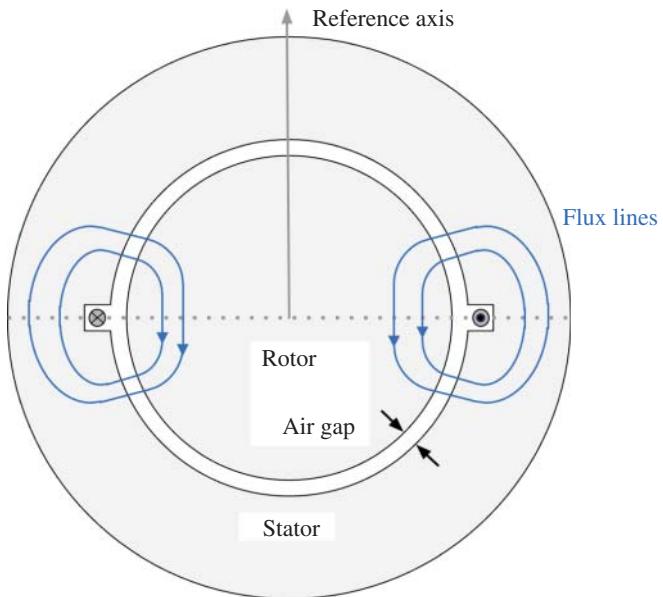


Figure 10.26 Simplified single-phase round rotor machine, with notional magnetic flux lines from the armature windings.

When a load is connected to the generator, it completes the armature winding circuit and allows current to flow in it. The flux associated with this load current is the *armature reaction*. The flux from all armature windings can be superimposed on the previous flux due to the field current, to determine the total flux across the air gap. This total air gap flux—or rather, its rate of change—then determines the *air gap voltage*, or the actual generated voltage that would be measured across the generator windings in the presence of load current (minus some small adjustments that we will come to).³² The flux due to the armature current surrounds the armature winding, not exactly in a circular shape, but also crosses the air gap in the radial direction.

The width of the air gap plays an interesting role: it effectively amplifies the flux density by being narrow. This makes sense in terms of a magnetic circuit, where the flux behaves like a “substance” moved by a magnetomotive force (see Section 2.6). The iron parts of the rotor and stator have a very high magnetic permeability (say, on the order of 1000 times that of air), so that they basically short-circuit the flux; in other words, the iron will not sustain much drop in the *mmf* (analogous to a conductor not sustaining an internal electric field or voltage). Consequently, on the path around the armature winding, the “drop” of magnetic force is concentrated across two gap crossings. The effect strengthens the magnetic field or flux density across the gap, inversely with width. This supports the intuitive sense that in order to achieve an effective transfer of power across the generator, we would like the air gap to be as narrow as possible.

The significance of the air gap also explains why the analysis of flux gets more complicated if the width is not uniform, as in the case of a salient pole rotor. That complication would be handled by viewing the whole system in a rotating reference frame, in which we can consider two components

concentrated in a single slot. The voltage across a realistic, distributed winding will in fact be a series addition of slightly out-of-phase contributions from each turn, but we need not worry about that subtlety here.

³² Applying superposition hinges on the assumption that the magnetic circuit representing the machine is *linear*, which makes the induced voltages additive.

of flux, one in line with the rotor axis called *direct*, and another at 90° to it called *quadrature* (see Section 4.3). For our simple round rotor, there is only a direct component.

While the rotor flux has a constant magnitude in time but rotates in space, the armature flux is fixed in space but varies sinusoidally in time. The phenomenal result for the three-phase reaction, as described previously (Figure 10.9), is that it appears as though it were constant in magnitude, but rotating in space.

A key finding, which can be derived from a Gaussian surface integral and the superposition of magnetic fields from three phases, is that the armature flux linkage interacting with a single phase as a function of time, $\lambda_{\text{ar}}(t)$, can be written as a constant times the current in only that phase:

$$\lambda_{\text{ar}}(t) = \text{const} \cdot i_a(t)$$

The presence of the other phases only changes the value of the constant, which is symmetrical for all three phases.

Where have we seen an expression like this before? Crucially, this proportionality has the same format as the equation for inductance, where L is the proportionality constant between flux and current, $\lambda(t) = Li(t)$ (Eq. (3.10) in Section 3.3.1, where we use symbol Φ for flux). This leads to an incredibly convenient interpretation: we may consider the constant a *fictitious inductance* and treat it as if it were a circuit element. This fictitious inductance is called the *synchronous inductance* and usually written L_s , but we'll use L_f :

$$\lambda_{\text{ar}}(t) = L_f i_a(t)$$

Keep in mind that this does not imply an actual physical circuit anywhere with such an inductance. Rather, L_f encapsulates various aspects of the generator's geometry: its length and radius, the width of the air gap, μ_0 , the number of turns in the armature windings, and a factor of π somewhere—the combination of which happens to describe the ratio of phase current to magnetic flux at every instant. A larger value of L_f corresponds to a machine with a stronger magnetic coupling.

10.7.1 Equivalent Circuit Model

We can now use that relationship for the purpose of creating a model, by superimposing the two sources of flux. The air gap flux λ_{ag} can be written as

$$\lambda_{\text{ag}} = \lambda_{\text{aa}'} + \lambda_{\text{ar}}$$

where $\lambda_{\text{aa}'}$ is the flux linking the winding aa' due to the rotor field, and λ_{ar} is the armature reaction. The air gap voltage v_{ag} is the rate of change of the air gap flux:

$$v_{\text{ag}} = -\frac{d\lambda_{\text{ag}}}{dt} = -\frac{d\lambda_{\text{aa}'}}{dt} - \frac{d\lambda_{\text{ar}}}{dt}$$

The first term is the open-circuit voltage across Phase a ,

$$\frac{d\lambda_{\text{aa}'}}{dt} = e_a$$

The second term can be written in terms of the fictitious inductance L_f and the current:

$$\frac{d\lambda_{\text{ar}}}{dt} = L_f \frac{di_a}{dt}$$

We can now incorporate the fictitious inductance into an equivalent circuit that represents the relationship between the generator internal voltage e due to the rotor field and the terminal voltage

v_a . The terminal voltage is almost but not quite equal to the actual generated voltage v_g , since we must subtract a small voltage drop due to series resistance r and leakage reactance X_l . The latter quantities are analogous to the series resistance and leakage reactance described in the nonideal transformer model; they should amount to less than 1% for the resistance and perhaps 10% for the leakage reactance.

The total *synchronous reactance* X_s combines the effect of the fictitious inductance $X_f = \omega L_f$ with the leakage reactance X_l as if they were circuit elements in series, such that $X_s = X_f + X_l$.

Using phasor notation, we can write an expression for the voltage V_a observed at the generator terminal under load:

$$V_a = E_a - rI_a - jX_s I_a \quad (10.1)$$

where E_a is the generator internal voltage on Phase a , I_a is the load current, and r is the winding resistance.

The relationship in Eq. (10.1) is indistinguishable from that produced by the equivalent circuit in Figure 10.27, which combines the fictitious with the actual electrical elements. Specifically, it elucidates how the generator terminal voltage V_a depends on the load current I_a . It also suggests a phase shift between the terminal and the internal voltage. Finally, by describing a physical process with mechanical and magnetic forces as if it were an electrical circuit, this equivalent circuit describes allows us to state the complex power transferred across the generator, and indicates theoretical limits of that power transfer.

The phasor diagram in Figure 10.28 illustrates the situation. This entire process is analogous to power transfer across a transmission line, where the generator internal voltage E_a corresponds to the sending end and the terminal voltage V_a to the receiving end voltage. Also like a transmission line (in fact, even more pronounced), the modeled generator impedance has $X_s \gg r$. Because that impedance cannot be zero, it imposes a theoretical limit for how much power can be transferred to the load (Sections 7.3.2 and 13.5), which is distinct from the thermal I^2R limit due to waste heat in the windings.

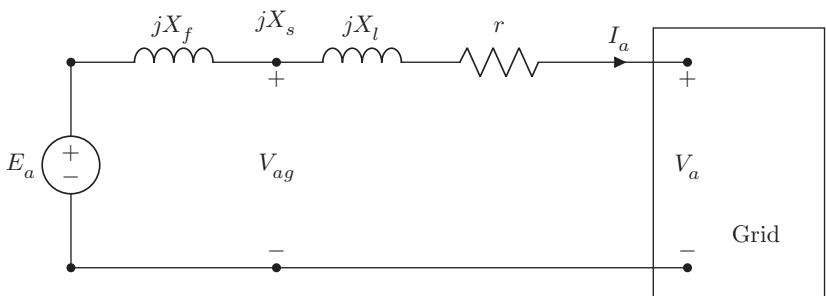


Figure 10.27 Simple equivalent generator circuit model.

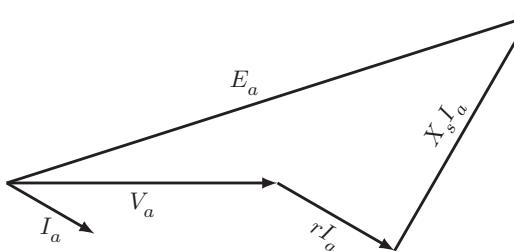


Figure 10.28 Phasor diagram for the equivalent circuit in Figure 10.27.

In fact, although we choose to isolate the generator from the transmission network for purposes of modeling, the generator windings, transmission lines and transformers in between all form a collective series impedance between the electromotive force and the load. Which impedance we are counting when we speak of transferring power across that impedance depends on which point on the circuit (such as the generator bus) we choose to designate as the reference.

10.7.2 Over- and Underexcitation

Remember that the load current I_a , including its magnitude and angle, is determined by the load. The phasor diagram in Figure 10.28 is a more rigorous way to support earlier assertions in Section 10.4.2 about the effect of reactive power demanded by the load. An increasingly lagging current causes a greater magnitude difference between $|E_a|$ and $|V_a|$. For a given amount of excitation or rotor field, this means that the terminal voltage $|V_a|$ will drop due to a more lagging current. Alternatively, in order to maintain a constant terminal voltage as reactive load increases, one can increase the rotor field current to raise $|E_a|$.

The term *overexcited* describes a generator whose internal voltage or excitation is greater than the terminal voltage ($|E_a| > |V_a|$), which means it is “generating” reactive power and supplying an inductive load with lagging current. Conversely, an *underexcited* generator is supplying a capacitive load with a leading current, and said to “take in VARs.” By rotating the current I_a and the voltage drop $jX_s I_a$ in the diagram so that I_a leads V_a , we can see that this results in $|E_a| < |V_a|$, analogous to a voltage rise on a transmission line.

The terminology of over- and underexcited is useful, because the terms “lagging” and “leading” power factor can get very confusing if we consider that a generator can also function as a motor, and *vice versa*. Recall that the load and generator conventions define positive current as entering or leaving the machine, respectively (see Figure 3.21). One consequence of the opposite current reference direction is that the analogous phasor addition to Figure 10.28 for a synchronous motor shows the voltage drop between E_a and V_a in the opposite direction than for a generator, with $V_a = E_a + I_a R + jI_a X_s$. The other consequence is that a 90° lagging current becomes a 90° leading current in the other reference direction. In the *generator convention*, a generator *producing* VARs is still said to operate at a *lagging* power factor, like a motor supplied by that same lagging current.

It is much less confusing to say that the generator is overexcited, and the motor is underexcited. Regardless of whether a machine is producing or consuming real power (i.e., acting as a generator or motor), and regardless of which convention is used for labeling the current direction, it always holds that *an overexcited machine produces* VARs and *an underexcited machine consumes* VARs.

10.7.3 Power Transfer

The power supplied by the generator to the load is given by the product of terminal voltage and current. Here we take another important conceptual leap: Although the current is physically determined by the load impedance, since the generator and the load share the same terminals and the same current, that power can also be described in terms of the internal generator properties which govern the relationship between terminal voltage and current. In other words, we can use the fictitious quantity X_s in the expression for power transferred. Instead of considering the current as the ratio of terminal-to-ground voltage to load impedance (which it also is), we write current as the ratio of the internal voltage drop (between excitation and terminal voltage) and the fictitious internal impedance. Writing complex power as $S = VI^*$, we obtain the analogous expressions for real and reactive power as derived for power transfer across a transmission line in Eq. (13.7).

From the transmission line context, we recognize the importance of the voltage phase angle difference δ : Assuming that $X \gg R$ for the line across which power is to be transmitted, real power transfer is proportional to $\sin \delta$. Substituting the generator's fictitious internal reactance, we see that a voltage phase angle difference δ is required between the excitation or internal generator voltage E_a and the terminal voltage. This has a very important implication. Note that the internal voltage E_a is directly in line with the physical rotor position, which can be mapped in degrees of rotation in either space or time, while the terminal voltage V_a lags behind it. The angle δ between E_a and V_a relates directly to the amount of real power injected into the network by the generator.

Now, which of these quantities is controllable? In a large power system, we generally assume that the frequency and phase of the terminal voltage V_a is beyond the individual generator's control; rather, it is determined by the interaction of many resources and loads across a large network.³³ In reality, especially if the generator is large compared to the rest of the system, it will somewhat affect V_a relative to a reference angle; we try to develop more intuition on this in Chapter 12. But when modeling a generator, it is beyond scope to account for such complexity of interactions. To keep the math tractable, we choose among three options: an *infinite bus*; a microgrid in which the generator is the only source driving the voltage; or a two-machine system.

Using the infinite bus assumption and taking V_a as externally given, we see that by pushing harder on the rotor—that is, by driving it forward relative to the alternating voltage at its terminal—we are advancing E_a relative to V_a , which increases δ and accordingly forces more electrical power into the grid. This also means that the grid will push back harder (with torque on the rotor that holds back E_a), and thereby act to stabilize δ , as discussed in Section 13.4. This is a fundamental fact that underlies the ability to operate many generators together synchronously.

The physical operating limits of the generator can be visualized in terms of how far the phasor diagram in Figure 10.28 can be stretched or distorted, while its connectivity is not negotiable. The magnitude of E_a is limited by the available rotor current—either based on what the exciter can supply, or what is safe without overheating the rotor windings. This constrains the reactive power $Q_a = |V_a||I_a| \sin(\angle V_a - \angle I_a)$ that can be delivered at the bus.

The angle δ is fundamentally limited by how hard the prime mover can push, but there is also a practical limit for the associated magnitude of I_a which limits the length of the phasor jI_aX_s . There is nothing in our model that constrains the possible values of I_a , so the armature current could, in theory, increase indefinitely. Realistically, though, someone would put a deliberate stop to the exercise to avoid overheating the generator windings.

Also, note that the real power $P_a = |V_a||I_a| \cos(\angle V_a - \angle I_a)$ (i.e., the component of armature current that is in phase with the bus voltage) has a theoretical maximum at $\delta = 90^\circ$. If we attempted to push the generator past this point, the current would have to grow huge but the grid would push back less, and the generator would lose synchronicity. In fact, Section 13.5 shows how the entire impedance through which the generator is supplying the load, including the transmission lines, will act to constrain power transfer, and Section 13.4 tells us not to intentionally operate anywhere near that theoretical limit.

To map these generator limits for real and reactive power, it is conventional to flip the axes to make P appear in the horizontal and Q in the vertical direction, consistent with the power triangle in the complex plane (even though this requires a mirror reflection of the phasor diagram, with

³³ A small generator connected to a larger grid can actually get damaged if it tries to regulate its own terminal voltage, especially if the grid voltage varies significantly, as the AVR makes futile attempts to compensate by adjusting the excitation current. In such a case, the generator needs to simply stick with a power factor and not worry about the terminal voltage.

V_a pointing upward). Together, these constraints help create the envelope of the *reactive capability curve* in Figure 10.23, which gives the practical operating constraints based on the prime mover power and generator internal heating limits.

Problems and Questions

- 10.1** A hydroelectric turbine can sustain a rotational speed of 225 rpm. If this is paired with a 60-Hz synchronous generator, how many poles should the generator have?
- 10.2** A junior operator in training is on his first day at a 100-MW steam generation plant. While the synchronous generator output is 90 MW, 40 MVAR, an alarm indicates that the bus voltage is down to 0.90 p.u. The junior operator asks you if the way to raise the voltage to 1.0 p.u. is to open the steam valve, so as to bring the output power of the unit to 100 MW. How do you respond?
- 10.3** The manufacturer of a four-pole induction motor rated 2 hp, 60 Hz specifies that it has 3% slip at full load.
- Approximately at what rpm does this motor operate when the load is 0, 1, and 2 hp?
 - Suppose the same machine is being used as a generator. What rpm would you expect when the mechanical power input to the generator is 1 hp?
- 10.4** A three-phase induction motor has the following information on its nameplate: HP 150, RPM 1785, AMPS 163, VOLTS 460, GUARANTEED EFFICIENCY 95.8, POWER FACTOR 0.89.
- Verify that this information is internally consistent by deriving the mechanical output in horsepower from the electrical input and efficiency. Note that the supply voltage is labeled as phase-to-phase but that the three motor windings are actually connected between each phase and ground.
 - What can you infer about the percentage of slip for this motor?
- 10.5** Consider the synchronous generator represented by the phasor diagram in Figure 10.28.
- In the operating condition depicted in the figure, is the generator over- or under-excited?
 - Sketch a phasor diagram for the case where the load supplied at the interface with the grid has unity power factor. How does this affect the excitation current?
 - Suppose the machine is reversed to operate as a synchronous motor, at unity power factor. Sketch a phasor diagram. (Hint: consider the sign of current in relation to voltage.)
 - Do you expect it's possible to operate a synchronous motor at different power factors? How would one accomplish this in practice?
- 10.6** Consider a generator with a round rotor. The terminal voltage is $V_a = 1.00$ p.u., the synchronous reactance is $X_s = 1.50$ p.u., the winding resistance is $r = 0.005$ p.u., and the current is $I_a = 1.00 \angle -45^\circ$ p.u.
- Draw a phasor diagram and find the internal generator voltage E_a to three significant figures.
 - Is r negligible in this situation?

- 10.7** A three-phase round-rotor synchronous generator, rated 16 kV (line-line) and 200 MVA, has a synchronous reactance of 1.60 p.u. and negligible resistance. It is connected to an infinite bus with a fixed voltage of 15 kV. The internal *emf* E is 24 kV (line-line) with a power angle $\delta = 25^\circ$.
- Find the per-phase line current I_a , in per-unit and in kA. (Hint: The base current should be 7.217 kA.)
 - Find the real and reactive power delivered by the generator.
 - Suppose the power input and excitation are adjusted so that the line current magnitude drops by 25%, while keeping the same power factor. Find the new internal *emf* E and δ for this operating condition.
 - The excitation is further adjusted so that the generator is delivering only real power at p.f. = 1.0, but at the same line current magnitude as in (c). Find the new internal *emf* E and δ for this operating condition.
 - Sketch a phasor diagram for the three operating conditions.

11

Matching Generation and Load

11.1 Load Frequency Control

In Section 10.4.1 on operational control of generators, the term *governor* was introduced, with a focus on what happens inside the machine and the power plant in response to changing load. Here, we take the perspective of the larger grid.

The key concept in load frequency control (LFC), also called *frequency regulation*, is synchronicity: namely that the shared, systemwide a.c. frequency is the same everywhere across the entire synchronous network. In fact, the assumption of a common frequency is an approximation. This quasi-steady-state analysis does not consider the dynamic behavior of the system or its components on very short time scales or during significant disturbances. In such dynamic conditions, we would actually find generators “swinging” against each other, meaning that the local frequency—as defined by the instantaneous rate of change of the local voltage phase angle—might vary. That type of phenomenon is studied in the context of stability analysis (Section 13.4). By contrast, frequency regulation focuses on how generator controls, individually and in aggregate, will respond to moderate load changes over the course of seconds and minutes. In this context, we ignore any local displacement or oscillations due to sudden changes and assume that the system frequency always equilibrates to a common value.

The synchronous a.c. frequency is the critical measure of power balance throughout the interconnected system: when aggregate generation equals aggregate load (including losses), the frequency holds steady. It is very important here to distinguish the notion of constancy from the actual value of frequency. Power balance doesn’t mean the frequency is necessarily equal to its nominal value of 50 or 60 Hz; it just means that the frequency is neither increasing nor decreasing.

A useful analogy is the water level in a tub. If the rate of inflow equals the rate of outflow, the water level holds steady. An imbalance between inflow and outflow means the water level will rise or fall.

We will distinguish *primary*, *secondary* and *tertiary* frequency control. Primary regulation is about arresting any change in frequency by restoring power balance as quickly as possible and making the frequency hold steady again. Secondary and tertiary frequency regulation deal with gently returning that steady frequency to its nominal value.

Because load cannot be perfectly forecast, some number of dedicated generators in a system will be assigned responsibility for responding to changes in load with appropriate changes in generation. This *frequency regulation* service may be remunerated through an *ancillary services* tariff that doesn’t focus on net energy supplied, but on the responsiveness in terms of speed and ramp rates. Generators accomplish frequency regulation by monitoring their rotational

frequency—which we assume, in a slow-changing system, is the same as the system frequency everywhere—and responding to changes through their governor.

A crucial fact about legacy grids is that this type of LFC was feasible in the 19th and early 20th centuries in a completely analog manner, without any digital technology or telecommunications. The system frequency itself, by virtue of being observable everywhere, was the only shared information necessary to coordinate large numbers of machines over long distances. It is for this reason that electric grids have been operable at all, without any of the data or advanced controls that support reliable and economical grid operation today.

An important quantity of interest is how rapidly the synchronous frequency will rise or fall as a result of a given power imbalance. In the first seconds or fractions of a second after an imbalance occurs, before any deliberate control actions take effect, the frequency change will depend on the stored rotational kinetic energy among the ensemble of generators in the system, which in turn depends on *rotational inertia*. This inertia is essential to justify our assumption about there being no sudden changes of frequency.

11.1.1 Inertia

Rotational inertia is the property of a spinning object to continue spinning at the same speed, until it experiences a force (i.e., physical work) to slow it down or speed it up. It is closely related to *angular momentum*, which is the product of the *moment of inertia* (I or J) and the angular speed ω (lowercase omega). The moment of inertia is a measure of an object's mass, and how far away from the axis of rotation that mass resides. For example, a large bicycle wheel with heavy tires has a greater moment of inertia than a smaller, lighter wheel: it takes more work to get up to speed, but it is also more stable—that is, it has a stronger tendency to continue spinning in the same direction.¹ We don't need to derive inertia of turbine-generators from first principles here. However, we should appreciate the physical intuition that heavy, large-diameter steel objects spinning at high speeds contribute a tendency to resist changes in the a.c. grid frequency.

If there is an imbalance between power injected and consumed throughout the synchronous system, kinetic energy is either added to or subtracted from the ensemble of spinning masses, causing the a.c. frequency to speed up or slow down. The more inertia is in the system, the more work goes into (or comes out of) this change of speed—and consequently, the smaller the frequency change for a given power discrepancy. This process constitutes the first level of shock absorption when an imbalance occurs suddenly. Note that both generators and motors connected to the grid provide inertia.

Generator inertia is further discussed in Section 13.4 about generator stability. For the individual machine, it is useful to define the per-unit *generator inertia constant* H as the ratio of rotational kinetic energy stored at synchronous speed, to the rated power output of the machine. This ratio has dimensions of time, with typical values on the order of a few seconds. It gives an intuitive sense of how long it would take a machine to exhaust its rotational kinetic energy at full rated output if the prime mover behind it suddenly stopped pushing. (Of course, this scenario is entirely hypothetical, since the machine would immediately disconnect from the grid.) Realistically, passive inertia can't last nearly as long as the H constant before the system frequency would go outside the allowable tolerance. This is why a quick response on the part of other generators is needed to restore power balance.

¹ The moment of inertia can be calculated as the integral over each bit of an object's mass and its squared distance from the axis of rotation. Angular momentum, the vector quantity that keeps a spinning bicycle wheel upright, is the product of inertia and angular speed.

11.1.2 Primary Frequency Regulation

Primary frequency regulation is the first active control response to counteract a power imbalance, performed by a generator *governor* (introduced in Section 10.4.1). It is a perfect example of feedback control. Here, a direct measurement of rotational speed is fed back to the actuator—say, a steam valve or water valve—that acts to increase or decrease the prime mover power. A simple proportional controller, known as *droop control*, will open or close the valve in direct proportion to the speed discrepancy.² The more the governor senses that the machine is spinning too fast, the more it will throttle down the prime mover, and conversely, if it is too slow, it will increase prime mover power up to its limit.

Today, generator governors can be programmed digitally, but the most intuitive design remains the classic centrifugal governor, employed by the Watt steam engine back in the 18th century. As illustrated in Figure 11.1, this elegantly simple concept relies on the centrifugal force to pull apart a pair of *fly balls* connected by a spring, whose separation is translated by some set of mechanical levers into a valve position for the prime mover.

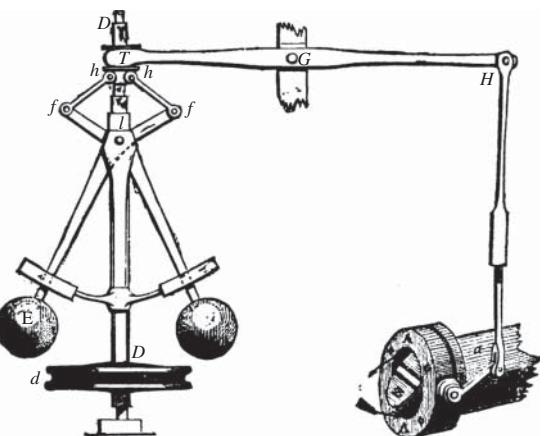
In this proportional, linear droop control, there are two parameters to adjust: the reference setting, and the slope. The reference setting is the desired valve position or power output when frequency is at its nominal value. The (negative) slope is called the *regulation constant R*. This corresponds to the ratio of frequency change to power output change, or

$$R = -\frac{\Delta f}{\Delta P}$$

Defined by its slope and intercept, the *droop curve* (or straight line, rather) is conventionally illustrated with frequency on the vertical and power on the horizontal axis, as in Figure 11.2.

A smaller absolute value of *R* corresponds to a more aggressive regulation effort, as a shallow slope implies a large change in power for a small change in frequency. It is intuitive to describe *R* in physical units of hertz per megawatt. However, the explicit value of *R* in Hz/MW will strongly depend on the size of the generator. It is common practice to “tune” different generators participating in frequency regulation to the same value of *R* in per-unit quantities (Section 8.7),

Figure 11.1 James Watt governor.
Source: Routledge/Wikipedia/CC BY 4.0.



² A proportional-integral controller would also take into account how long the discrepancy had been accumulating or losing energy, and a derivative term would take into account the acceleration. Modern governor systems can incorporate such advanced control strategies, but they are beyond the scope of the classical textbook treatment of frequency regulation, which assumes only linear control.

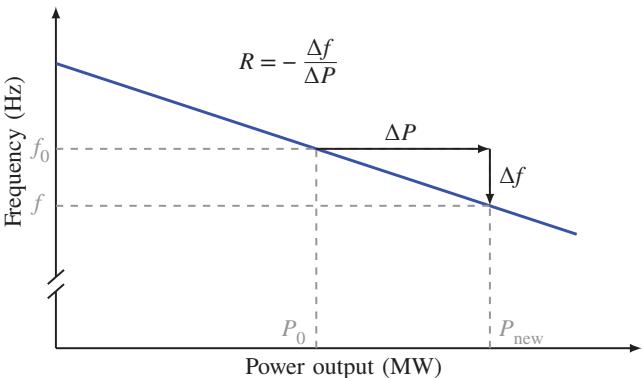


Figure 11.2 Droop curve showing the relationship between Δf and ΔP .

with the most typical values of R being 0.04 or 0.05 p.u. This practice allocates the contributions from different power plants in proportion to their capacity.

For example, a unit rated at 1000 MW (S_{BASE}) with a regulation constant of $R = 0.05$ p.u. at 60.00 Hz nominal (base) frequency would respond to a system frequency decrease of 0.01 Hz with a power increase of 3.33 MW:

$$\Delta P_{\text{p.u.}} = -\frac{\Delta f}{R} = -\frac{-0.01 \text{ Hz}/60 \text{ Hz}}{0.05 \text{ p.u.}} = 0.00333 \text{ p.u.}$$

$$\Delta P_{\text{MW}} = 0.00333 \text{ p.u.} \cdot 1000 \text{ MW} = 3.33 \text{ MW}$$

The extreme case of a horizontal line, with zero R , is called an *isochronous* machine. A generator in isochronous mode will do anything it can to keep the frequency exactly at its setpoint. This control mode is only used in stand-alone applications or microgrids, where a single generator is responsible for maintaining a.c. frequency. It is easy to see that more than one machine in isochronous mode on a synchronous grid would misbehave and fight each other, if there were the slightest mismatch in their set points.

The opposite case is a vertical line, which simply means that the machine is not performing any droop control at all. It is simply operating at a chosen power output, irrespective of frequency. When a generator is running at full capacity with no droop, it is said to operate *on the load limit*, as opposed to *on the governor*.

Primary frequency regulation refers to movement up and down the droop curve. As grid frequency varies—and with it the synchronous machine speed, which we assume to be locked in step—the governor adjusts the prime mover power and thereby helps counteract the variations. Again, it is important to emphasize that this control action requires only local sensing with no external communication and can be executed instantly, except for any lag introduced by mechanical components and steam flow.

When a change in system load occurs, this will physically result in an increasing or decreasing system frequency, as kinetic energy is gained or lost. In *primary* response to the changing frequency, the amount of power generated will be automatically adjusted by the governor to equal that of the new load (assuming, of course, that there is enough capacity available and generators don't run into their limits). When generation and load are again equal, the frequency holds steady—but it will no longer be at its original value.

11.1.3 Secondary Frequency Regulation

Secondary frequency response involves shifting the generator set points in order to return the system to its nominal frequency. This means shifting the droop curve up or down (keeping the same slope, but changing the intercept) so that the amount of power generated aligns with the desired frequency. Because there is no immediate harm in operating at a constant but slightly off-nominal frequency, secondary frequency regulation can take tens of seconds or even minutes to complete. The process is best described by example.

Example

Consider the standard droop curve in Figure 11.3. Suppose it represents a generator rated at 1000 MW in a 60 Hz system and that the slope is $-R = -0.05$ p.u. This corresponds to

$$-R = -0.05 \text{ p.u.} \frac{60 \text{ Hz}}{1000 \text{ MW}} = 0.003 \text{ Hz/MW}$$

Suppose the generator is initially operating at 600 MW when the system frequency is $f_0 = 60.00$ Hz. Now, due to a disturbance somewhere (say, a sudden increase in load), the system frequency drops by 0.15 Hz down to 59.85 Hz, a change of

$$\Delta f = -0.15 \text{ Hz} = \frac{-0.15 \text{ Hz}}{60 \text{ Hz}} = -0.0025 \text{ p.u.}$$

Sensing the change in frequency, the governor responds by increasing power output, by an amount

$$\Delta P = -\frac{\Delta f}{R} = -\frac{-0.0025}{0.05} = 0.05 \text{ p.u.} = 50 \text{ MW}$$

so that the new operating point is 650 MW. The change corresponds to movement along the droop curve in Figure 11.3, from the “initial state” to the point labeled “primary response.”

Let’s assume the frequency holds steady in this new condition, meaning that power in the grid is balanced, since our generator has made up for the shortfall. Note that from the vantage point of an individual generator, it is impossible to observe directly how big a load change occurred, or how many other units contributed to the response. In any case, by observing that the frequency holds steady, we know that a crisis has been averted for now. In principle, the grid could continue to operate in this state without any further control action.

However, the new frequency is not really where we want to operate, for both physical and economic reasons to be discussed below. If the additional 50 MW of load is here to stay, we want to

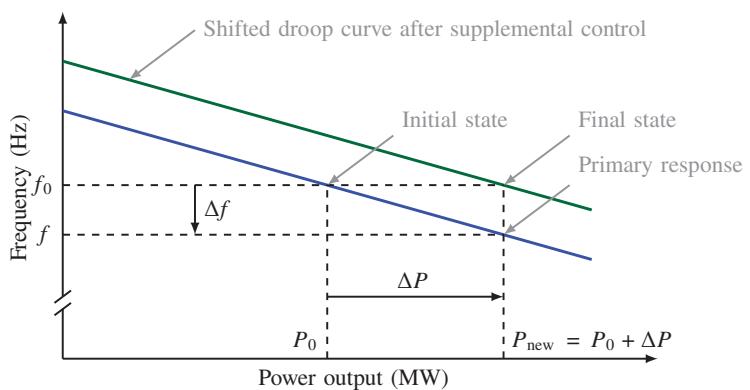


Figure 11.3 Droop control with a single generator.

adjust the control settings so that we return to the nominal frequency of 60.0 Hz with the new load. To accomplish this, we need to shift the droop curve upward so that it goes through the operating point of 650 MW and 60.0 Hz (labeled “final state” in Figure 11.3). This is called *supplemental*, *supplementary*, or *secondary* frequency control. In the centrifugal governor, it would amount to adjusting the lever position on the steam valve for a given separation of the fly balls.

11.1.3.1 Multiple Generators

If multiple generators in a synchronous system are performing frequency regulation, we assume that they are all observing the same frequency. Thus, their operating points at any given instant will have the same vertical coordinates. We can then add their droop curves horizontally to find the total, systemwide change in power output corresponding to a given change in frequency, or *vice versa*.

The sum of responses from all the synchronous generators in an area is called the *area frequency response characteristic*, denoted by β , in per-unit or in MW/Hz, often scaled to MW/0.1 Hz. Beta is just the sum of inverse regulation constants R_i , or, for a set of N generators:

$$\beta = \frac{1}{R_1} + \frac{1}{R_2} + \frac{1}{R_3} + \cdots + \frac{1}{R_N}$$

With multiple generators, we can use β to find the relationship between a systemwide load and frequency change:

$$\Delta P_{\text{sys}} = -\beta \Delta f$$

The idea is simple; the caveat is making sure that the units are scaled correctly.

Figure 11.4 illustrates a case with two units jointly responding to the same frequency change on the system. The arithmetic is made a bit more interesting by the fact that the units have different power ratings and different regulation constants.

Example

In this example, Unit 1 is rated 400 MW with $R_1 = 0.04 \text{ p.u.} = 0.006 \text{ Hz/MW}$, and Unit 2 rated 600 MW with $R_2 = 0.05 \text{ p.u.} = 0.005 \text{ Hz/MW}$ (taking $f_{\text{BASE}} = 60 \text{ Hz}$, and the rated capacity of each generator as its S_{BASE}).

We are interested in the allocation of primary frequency response between the two units. By inspection of the above values, we already know that Unit 2 with the smaller R in Hz/MW will respond slightly more aggressively to a change in system frequency, with a bigger ΔP for a given Δf . This is due to its greater rated power in MW, in spite of its greater per-unit value of R .

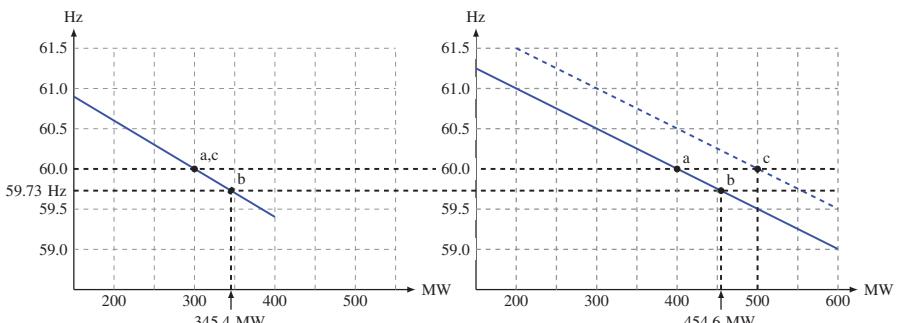


Figure 11.4 Droop control with two generators. Points a are the initial state, b after primary response, and c the final state after secondary response, in which only Unit 2 participates.

Let's work through a sample scenario. The initial operating points are 300 MW for Unit 1 and 400 MW for Unit 2, at 60.00 Hz, with a total system load of 700 MW. Suppose that now there is a sudden 100-MW load increase.³

We can choose to work in per-unit quantities or with Hz and MW; the latter is perhaps more intuitive but less elegant when applied to complicated situations. Either way, our first task is to identify the systemwide frequency change, since this is what both generators observe in common.

This problem is straightforward to answer graphically: we add the two droop curves horizontally and finding the vertical point (frequency) at which the combined horizontal value (power) has increased by 100 MW. Note that for determining the systemwide change in frequency, only the overall β matters, not how the frequency response contribution will be allocated among some number of generators in the system.

Algebraically, we can write

$$\Delta f = -\frac{\Delta P}{\beta} = -\frac{100 \text{ MW}}{\frac{1 \text{ MW}}{0.006 \text{ Hz}} + \frac{1 \text{ MW}}{0.005 \text{ Hz}}} = -0.273 \text{ Hz}$$

or equivalently:

$$\Delta f_{\text{p.u.}} = -\frac{100 \text{ MW}}{\frac{400 \text{ MW}}{0.04 \text{ p.u.}} + \frac{600 \text{ MW}}{0.05 \text{ p.u.}}} = -0.00455 \text{ p.u.}$$

The new system frequency is $60 - 0.273 = 59.727$ Hz, or $1.00 - 0.00455 = 0.995$ p.u.

Now we can solve for the response from each generator, using quantities either in hertz or per-unit:

$$\Delta P_1 = -\frac{-0.273 \text{ Hz}}{0.006 \text{ Hz/MW}} = 45.4 \text{ MW}$$

$$\Delta P_2 = -\frac{-0.00455 \text{ p.u.}}{0.05 \text{ p.u.}} = 0.091 \text{ p.u.} = 54.6 \text{ MW}$$

As expected, the sum of the responses equals the initial load change of 100 MW, with Unit 2 pulling slightly more than half the weight. Calamity has been averted, and the system is operating at a steady but undesirable new frequency.

Let's suppose that only Unit 2 performs supplemental control to drive the frequency back to 60.0 Hz. This amounts to shifting its droop curve such that it will absorb the entire 100-MW change by itself. As Unit 2 gradually increases its power output and the frequency inches upward, Unit 1, which continues to perform primary frequency regulation at all times, will respond by moving back along its fixed droop curve and gradually reduce power output. The detailed dynamics of this control process are beyond the scope of our analysis. Intuitively, we can appreciate that an abrupt shift might cause an oscillation, as all generators on the system adjust and might overcompensate, so the changes need to be made gently.

In the final state, Unit 1 will have returned to its original operating point as if nothing had happened. However, both generators have performed an important service to the system: Without the help of Unit 1, Unit 2 would have had to make a much more drastic change in power output to absorb the initial load change—possibly beyond its physical ability, and possibly causing reverberations throughout the grid.

The participation of many generators in primary frequency control buys time for those tasked with secondary frequency control to do their job. Generators may participate in providing one or

³ This is quite a dramatic disturbance for such a small system, but it serves to illustrate our case. Whether these generators would be able to smoothly absorb that shock is outside the scope of the present analysis.

both services, and get paid for their response according to ancillary services tariffs. When many generators share the task of secondary or supplemental frequency regulation, a quick calculation is performed by a coordinating entity to allocate the megawatt response among the available resources. This calculation takes into account their respective costs and stated capabilities. For example, a given generator may have pledged to provide regulation services within some range of power output, or constrained by some *ramp rate* in megawatts per minute. Especially for steam generation, thermal inertia limits the speed at which power output can be varied (since it takes time to build up steam pressure). Secondary frequency control also takes into consideration transfers between neighboring areas; this is discussed in Section 11.1.5 on the *area control error*.

Historically, updated set points for each generator were communicated to operators by telephone—underscoring the idea that once primary frequency control resources have responded, the need to make adjustments is not terribly urgent and can be made on a timescale commensurate with human action. However, faster response usually means better optimization. Today's standard for secondary frequency response is an *automatic generation control* (AGC) signal transmitted directly to the participating generator governors, with updates as often as every few seconds. On a slower time scale of minutes, *tertiary* frequency control performs a more thorough economic optimization to make bigger adjustments when needed. The decision process concerning how much power to recruit from each generator is discussed in Section 11.2 on Economic Dispatch.

In summary, the roles of primary, secondary and tertiary frequency control are as follows: primary stops the frequency from changing; secondary returns the frequency to the desired value; and tertiary minimizes cost.

11.1.4 Frequency Tolerance

Why is it important to return to the nominal frequency after any disturbance? One historical reason discussed in Section 5.2 is timekeeping with analog clocks. Before the proliferation of electronics and various sources of accurate time signals, from oscillating crystals to GPS satellites, clocks plugged into a.c. grid power served as the authoritative source of official time. Even in the days before automated stock trading, where milliseconds can mean millions of dollars, there was some economic value for industrial societies in having a reasonably accurate shared reference time. But timekeeping is not the most serious problem.

The major physical reason for managing frequency within tight bounds is rooted in the fact that at off-nominal frequencies, rotating machines might experience excessive mechanical forces they were not designed to withstand. The amount of over- or underfrequency considered dangerous to machines of various types is a wide gray area. Ultimately, someone has to weigh the trade-off between the risk of damaging expensive equipment on the one hand, and the risk of interrupting loads on the other hand, to arrive at an engineering judgment about where to set reasonable limits for the system. These limits are expressed in the trip settings of over- and underfrequency relays. When a relay trips, it means that some part of the system is disconnected, and some customers may lose power (see Section 7.5).

Given these relay settings, there is reason not to wait too long to recruit supplemental regulation to restore frequency: if the frequency is held at a value significantly above or below nominal, even if it is stable there, the system is operating closer to the threshold and thus more vulnerable to having a relay tripped by the next disturbance.

Different system operators or jurisdictions can make different choices regarding how aggressively they act to regulate frequency, and where they set their protection thresholds. However, because interconnected areas are always physically interdependent, rules and standards are negotiated to make sure the control actions in one area don't cause adverse consequences for their neighbors.

This brings us to another reason for managing frequency within a tight tolerance band: managing power exchange between and among control areas. Simply put, the more the frequency wanders, the harder it will be to keep track of power imports and exports, who is doing their job of frequency regulation, and how much money needs to change hands across jurisdictions.

11.1.5 Area Control Error

Trading electricity across borders through a shared transmission network poses some unique challenges to the accountant. Unlike most other goods—a truckload of tomatoes, or a railcar of coal—electricity is not a tangible object that can readily be tracked. Instead of deliberately driving electrons from seller to buyer, electric power markets must submit to Kirchhoff's laws and settle the bill after the fact. In addition to maintaining the physical power balance throughout an interconnected system, maintaining the desired balance of business transactions involves a continual process of approximations and adjustments. Because actual load varies in real time with some unpredictable noise, and because contingencies happen, no advance planning effort can perfectly control or predict the amount of power that will flow on any given transmission line at any given instant.

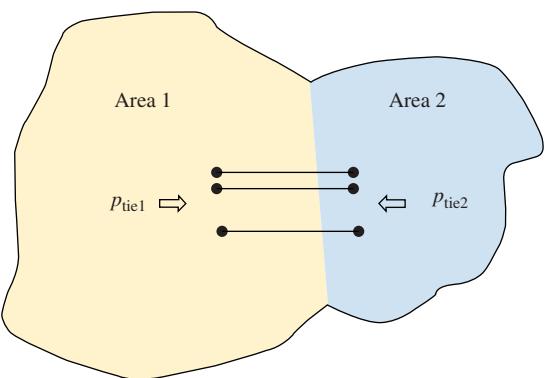
Special attention is paid here to *tie lines* that connect different jurisdictions called *balancing authorities* in the United States, also sometimes referred to as *operating areas*. A balancing authority is an organizational entity responsible for reconciling power supply and demand within a specific geographic region, including the oversight of sales transactions and management of ancillary services. For our purposes here, we ignore the vast complexities of how balancing authorities actually accomplish this. In the simplest terms, their job is to keep track of the amount of generation and load within their area, and how the difference is being met with imports or exports.

Balancing authorities can be large or small—say, encompassing a city, state or country—and they may border on one or several neighboring balancing authorities. In each case, there will be a finite number of transmission lines that cross the borders; these are the tie lines, as illustrated in Figure 11.5.

For each time step (say, hourly or 15 minutes), balancing authorities within a larger interconnected region will compare notes on their anticipated load and their internal generation schedule, along with sales transactions spanning other areas. A power flow analysis (see Chapter 12) then reveals the expected flow on each tie line resulting from the superposition of all these actions. For any area, total tie line flows out of the area must (by energy conservation) equal the sum of all generation, minus total loads and losses inside the area.

We don't expect tie line flows to ever be zero, since individual areas are rarely self-sufficient (if they were, it would have made no sense to build the transmission capacity in the first place).

Figure 11.5 Visualizing tie line flow between neighboring control areas.



The point, rather, is that based on the forecast load and all the scheduled transactions, a *scheduled* set of tie line flows is calculated. If the actual tie line flows in real time match the schedule, this means that all transactions have been perfectly executed and accounted for, and there is zero error.

In reality, of course, there will be discrepancies. Usually, they are small variations because neither load nor generation can be perfectly forecast. Bigger changes may be due to contingencies like the unexpected loss of a generator, or a transmission line outage that impacts power flow on all the other lines. Even if the total generation in each area remains unchanged, the total transmission losses in the system can vary and impact tie line flows (e.g., if load is shifted between generators from opposite geographical corners of an area). Also, generators from all areas will respond with primary frequency regulation when changes occur in any one area, shifting the balance temporarily. Therefore, tie line flows are continually monitored, and a *tie line flow error* Δp_{tie} is defined as the difference between actual and scheduled flow:

$$\Delta p_{\text{tie}} = p_{\text{tie}} - p_{\text{sched}}$$

We are now ready to revisit the question of how to choose new generator set points for secondary frequency regulation, to be communicated by AGC. Basically, AGC chases a quantity called the *Area Control Error* (ACE). It is defined such that when ACE = 0, all is well.

The ACE has two components: frequency error, and tie line error. Each balancing authority observes its own ACE, which varies from moment to moment. The units of ACE are megawatts. When the number is positive, it means that there is excess generation; when it is negative, it means that there is a lack of generation in the area.

$$\text{ACE} = \Delta p_{\text{tie}} - B_f \Delta f \quad (11.1)$$

The frequency error term captures the *systemwide* excess or lack of generation. It is given by the frequency deviation Δf in hertz, multiplied by the *frequency bias constant* B_f .⁴ The frequency bias constant B_f has the same units as the area frequency response characteristic β and otherwise resembles it; in fact, B_f is often assigned the same per-unit value as β . However, they are distinct quantities and need not be equal. As the sum of inverse regulation constants, β reflects the physical behavior of *primary* frequency response, whereas B_f sets the policy for *secondary* response in a given area. A larger value of B_f indicates a more aggressive response to return frequency to its nominal value.

We adopt the sign convention here where B_f is a negative number, which can be confusing (and it is not how we treated β). Some references treat B_f as positive and change minus to plus in Eq. (11.1). The consistent idea regardless of convention is that when frequency is too low, the $B_f \Delta f$ term tends to make the ACE negative, indicating that generation should be increased (and *vice versa*).

The tie line flow error term Δp_{tie} in the ACE captures the discrepancy between and among areas. Such discrepancy is inevitable because any given disturbance or deviation from planned generation and load will occur in one area or another, meaning that the actual tie line flows in real time will deviate from what was scheduled.

The crucial property of the ACE is that if every area successfully regulates its internal resources to return its own ACE to zero, then any discrepancy will have been addressed by the area in which it occurred. For example, if a generator is suddenly lost in one area, all its neighbors will provide primary frequency response to help stabilize the system. However, the neighbors will then observe

⁴ Let us re-emphasize that in the present scope of analysis, we always assume that frequency is in a quasi-steady state. This means that although we expect frequency to change from one time step in our analysis to the next, we assume that the frequency has equilibrated to a value which is *the same across all areas* for any given moment at which we analyze the system. In other words, we do not consider any oscillations or uncontrolled dynamic responses to change. Even if we can measure transient local discrepancies in frequency, that information would not be actionable within the context of load frequency control.

a positive ACE, as their generation resources are now compensating for power flow that originated in the area which experienced the loss, while that area will observe a negative ACE, as its export balance dropped. By restoring each ACE to zero, replacement generation resources will be recruited specifically within the area where the loss occurred. Because ACE combines the frequency and tie line flow errors, it will ensure that the neighbors don't back off too quickly on their support. For example, if the affected area cannot muster the resources to respond to its loss, the system frequency will remain too low and thereby prompt the neighbors to continue contributing more power than was scheduled. A numerical example is illustrated below.

Note that this control paradigm requires a very limited set of measurements, and minimal knowledge about the actual nature of events on the grid. Frequency is measured directly at each generator bus, and—since at this scale of analysis, it has the same value everywhere—does not need to be communicated. The only other sensor data required are the tie line flows, which are communicated to the location within each area that computes the ACE, and in turn issues AGC commands to specific resources signed up for performing secondary frequency regulation services. The key is that individual generators need not be aware of the type or cause of disturbance they are responding to: all they have to know is whether to increase or decrease power output at any moment.

Example

Figure 11.6 illustrates a situation with three neighboring areas that have different power bases and are using different droop settings and frequency bias constants. In the initial balanced condition, Area A is importing 1000 MW and Area C is importing 3000 MW from Area B, 1000 MW of which flow through Area A. Note that the tie line flows are contrived to have round numbers; they would be a function of the various line impedances throughout the network, and the information necessary to determine them is not given in this problem. A key concept illustrated by this scenario is that we care only about the *net* tie line flow for each area, which would be -1000 MW for Area A regardless of how the flows arrange themselves across the network. For simplicity and clarity, we ignore losses throughout this example.

Now suppose a disturbance occurs: Area A suddenly loses 475 MW of generation. What happens?

To determine the primary response, we must first find the frequency change Δf throughout the system. Here, we combine the droop curves from all three areas into a collective frequency

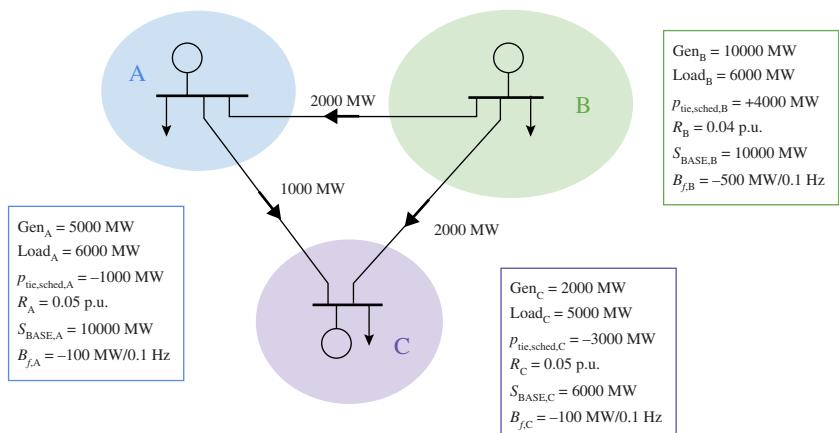


Figure 11.6 Three areas sharing frequency regulation resources (initial operating condition for text example).

response characteristic β . For simplicity, we assume the regulation constants are the same for all generators within a given area; we don't need to know how many distinct units the resource actually comprises. Note that the horizontal addition of droop curves works the same way inside or across areas, since the generator governors are physically oblivious to the jurisdictional boundaries. From the respective contributions to β we can see that Area B offers the most MW response per unit frequency change, since it has a lower R than Area A, while Area C offers the least since it has a smaller base capacity.

$$\begin{aligned}\Delta f &= -\frac{\Delta p}{\beta} = -\frac{475 \text{ MW}}{\frac{10,000 \text{ MW}}{0.05 \text{ p.u.}} + \frac{10,000 \text{ MW}}{0.04 \text{ p.u.}} + \frac{6000 \text{ MW}}{0.05 \text{ p.u.}}} \\ &= -\frac{475}{200,000 + 250,000 + 120,000} \text{ p.u.} = -0.0008333 \text{ p.u.}\end{aligned}$$

$$\Delta f = -0.000833 \text{ p.u.} \times 60 \text{ Hz} = -0.05 \text{ Hz}$$

The primary response from each area is now straightforward to find:

$$\begin{aligned}\Delta p_A &= -\frac{\Delta f}{R_A} = \frac{0.000833}{0.05} \text{ p.u.} = 0.0167 \text{ p.u.} = 167 \text{ MW} \\ \Delta p_B &= -\frac{\Delta f}{R_B} = \frac{0.000833}{0.04} \text{ p.u.} = 0.0208 \text{ p.u.} = 208 \text{ MW} \\ \Delta p_C &= -\frac{\Delta f}{R_C} = \frac{0.000833}{0.05} \text{ p.u.} = 0.0167 \text{ p.u.} = 100 \text{ MW}\end{aligned}$$

To appreciate the importance of the assistance from the other areas, consider what might have happened if Area A were isolated and had to carry the entire 475 MW load change by itself: the frequency change would have been $-475/200,000 = -0.002375$ p.u. or -0.14 Hz, low enough to risk tripping underfrequency relays and potentially cause a blackout. Instead, the day has been saved by stabilizing the frequency at $60.00 - 0.05 = 59.95$ Hz.

Next, how will each area adjust during supplemental frequency regulation? We use the ACE, which is given by the sum of the frequency and tie line flow errors.

For Area A, where the loss occurred, both terms of the ACE are negative, indicating that more generation is needed. Area A sees a tie line flow error from the contributions of both its neighbors, such that its net import is 1308 MW instead of the scheduled 1000 MW:

$$\begin{aligned}\text{ACE}_A &= \Delta p_{\text{tie},A} - B_{f,A} \Delta f = -308 \text{ MW} - (-100 \text{ MW}/0.1 \text{ Hz}) \times -0.05 \text{ Hz} \\ &= -308 - 50 \text{ MW} = -358 \text{ MW}\end{aligned}$$

Note that Area A sees generation still short by 358 MW, even after its initial correction of 167 MW during primary frequency control. The sum of the two seems like an overcorrection (since $358 + 167 > 475$), but as the frequency recovers, the ACE will gradually shrink. Area A's ACE will reach zero once the frequency has been fully restored and its generation has increased to absorb the full initial 475 MW shortfall.

Now consider the ACE for Area B. We know that it is producing an extra 208 MW. This assumes that its own loads and losses have not changed, so its tie line flows should now add to 4208 MW, up from the scheduled 4000 MW. For weighting the frequency error, we use the frequency bias constant $B_{f,B} = -500 \text{ MW}/0.1 \text{ Hz}$. Thus:

$$\begin{aligned}\text{ACE}_B &= \Delta p_{\text{tie},B} - B_{f,B} \Delta f = 208 \text{ MW} - (-500 \text{ MW}/0.1 \text{ Hz}) \times -0.05 \text{ Hz} \\ &= 208 - 250 \text{ MW} = -42 \text{ MW}\end{aligned}$$

At this moment, the low frequency dominates the ACE, prompting Area B to read that its generation is too low and provide still more support, despite the fact that it is already exporting more than scheduled. The rationale is to keep the lights on first, and worry about the accounting later. As the frequency gradually returns to nominal, the frequency error term in the ACE shrinks, and the tie line flow error term will come to dominate. We don't know the time steps of the control, but we can say that Area B will back off on its support once the frequency has recovered to the point where the frequency and tie line flow errors cancel.

Area C has $B_{f,C} = 100 \text{ MW/Hz}$. Its net tie line flows are reduced by the 100 MW it is contributing to Area A, yielding

$$\begin{aligned}\text{ACE}_C &= \Delta p_{\text{tie},C} - B_{f,C} \Delta f = 100 \text{ MW} - 100 \text{ MW/0.1 Hz} \times -0.05 \text{ Hz} \\ &= 100 - 50 \text{ MW} = +50 \text{ MW}\end{aligned}$$

Because Area C's frequency bias constant was set lower, it places a comparatively greater emphasis on maintaining its scheduled imports and exports, and will contribute less effort toward restoring frequency. This would make sense, for example, if Area C is small and has fewer resources to offer. After contributing to primary frequency control, its ACE is already positive, meaning that it will immediately begin to reduce its generation. The ACE should shrink in the process of performing secondary regulation and reach zero after the area has returned to its originally scheduled import and export levels.

The frequency bias constants and time steps for sending new AGC signals to participating generators are chosen to make transitions smooth, like steering a ship, and to minimize the risk of a widespread blackout. The entire process of recovering from a big step change will take at least a few minutes. In the meantime, it is quite possible that another disturbance will happen; at least small changes in load are bound to occur. Therefore, chasing the ACE is a continuous process that doesn't mark where one disturbance ends and the next one begins. In practice, grid operators would look at a screen that prominently displays the ACE next to the systemwide frequency, and gets updated at least every few seconds—like a heart rate monitor on a medical patient. When all is going well, the frequency will wander around the neighborhood of 60.00 (50.00) Hz, and the ACE will wander around in small numbers of MW compared to each area's overall load. Figure 11.7 illustrates the scale of normal frequency variations during a relatively uneventful 45-minute period.

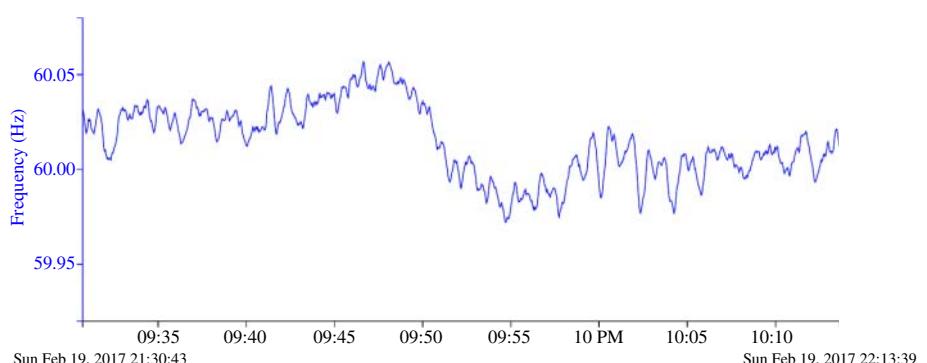


Figure 11.7 Grid frequency on a typical day in California, measured at a wall outlet with a micro-phasor measurement unit (PMU) and visualized in the Berkeley Tree Database (BTrDB).

11.2 Economic Dispatch

The process of determining how much power should be contributed at any given time from each one of the many generators in a large, interconnected power system has been the subject of much academic analysis as well as practical experimentation. Over recent decades, this process has evolved in two major ways: from vertical integration of the industry to competitive generation markets, and from self-contained optimization to much faster and data-rich online computation. Given the complexity and nuance of modern market processes, it seems almost ridiculously quaint to study the simple, classical formulation of the *economic dispatch* problem. We introduce it here because it is conceptually and pedagogically valuable.

Conceptually, the economic dispatch solution represents an idealized version of the outcome one hopes that competitive markets will achieve in practice: minimizing the overall cost of satisfying a particular level of demand. It also features a key insight of market economics: that in the configuration at the minimum total cost, the marginal costs for all suppliers are equal. The argument is simple yet profound. Suppose one supplier had a different marginal cost than another of providing one additional megawatt. Then we could reduce overall cost by having the cheaper supplier take over production of an extra megawatt from a more expensive supplier (assuming they are not up against their production limit). Therefore, the previous condition could not have been a minimum.

This expresses the general fact that at any extremum (minimum or maximum), the derivative (or gradient) of a function—that is, its rate of change with respect to the independent variable (or variables)—must be zero.⁵ In our case, the independent variables are the megawatt contributions from all the individual suppliers.

The solution of an idealized, toy version of economic dispatch is also instructive pedagogically as an illustration of the mathematical method of *Lagrangian multipliers*, which has many other applications in engineering. It may be valuable for students of power engineering to have some first-hand feel for the problem style in general, and the term “economic dispatch” in particular, even if our textbook approach bears little resemblance to modern real-world operations. Finally, the Lagrangian method is a classic example of how power systems analysis was once handled with pencil and slide rule as the only computational tools. It serves as a reminder that the many simplifications and approximations in this field arose out of sheer necessity.

11.2.1 Filling in the Load Duration Curve

Before we embark on the mathematical formulation, let us contemplate the big picture of the desired outcome. Recalling the *load duration curve* (LDC) from Section 6.4, successful economic dispatch over an entire year can be visualized as “filling in” the area under the curve with various types of generation so as to minimize overall cost, while meeting all operating constraints.

Traditionally, there are three general categories of generation: *baseload* generation units, which produce the cheapest energy and are best operated on a continuous basis (e.g., coal or nuclear plants); *load-following* units that respond to changes in demand (typically, hydroelectric and selected steam generation units); and *peaking* units that are expensive to operate and are used to meet demand peaks (e.g., gas turbines). More recently, nondispatchable solar and wind power, also described as *variable* or *intermittent* generation, has become another significant category.

⁵ In other words, the top of a hill and the bottom of a valley must be flat, at least in one spot. This is true for differentiable functions without sharp edges or spiky features. The notion of a gradient extends from the intuitive two-dimensional case of hills and valleys to any number of dimensions.

Before these resources accounted for a large fraction of supply, they were often just treated as negative load, combining their statistical variability with that of demand, and thus pushed outside of the problem formulation. Let us adopt that classical perspective here, for the sake of learning about a tractable math problem.

The area under the LDC corresponds to the total, systemwide annual energy demand. Note that since the vertical axis of the LDC measures units of power (MW) and the horizontal axis measures units of time (hours, where it does not matter that the hours do not occur sequentially), the area under the curve can be measured in units of energy (MW-h). Demand, the outer bound of the area, is exogenous to our problem: we assume nothing can be done to change it.⁶

To fill the area with available generation resources, we can imagine using markers of different widths to color or shade in Figure 11.8. Different types of units are scheduled with baseload as a priority, load-following plants added in as required, and peakers kept in reserve for the extreme days or hours. The energy output of baseload units throughout the year is represented by the rectangular area filling the lower portion of the curve, with the combined power output of baseload units remaining nearly constant. Load following and other units operated at variable power levels are shown together as filling in the central portion of the curve. Their power output actually varies from hour to hour and day to day, as can be seen on the left-hand portion of the diagram where time is shown sequentially; the LDC simplifies this temporal profile while highlighting overall energy. Finally, the contribution from peaking units is immediately recognizable as the area that fills the top of the curve.

This scheduling process is obviously idealized. In reality, generation units have specific constraints on their operation that must be taken into account. These constraints include scheduled outages for maintenance (and refueling, in the case of nuclear units), unscheduled outages, and limitations on the *ramp rates* at which particular units can safely increase and decrease their output power. For large thermal plants, the ramp rates may significantly affect scheduling. The combination of constraints on unit availability produces a continually changing menu of generation capacity throughout the system, from which the optimal contribution levels are to be determined for a given day and time.

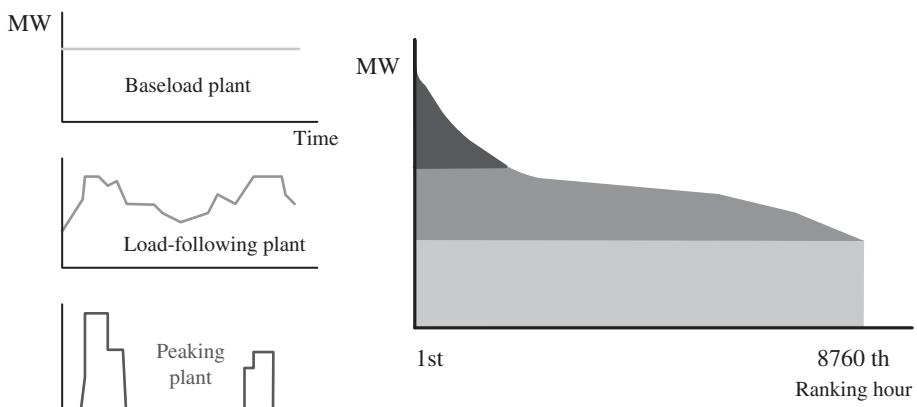


Figure 11.8 Generation scheduling with the load duration curve.

6 This is quite problematic from a modern standpoint, as it might cause one to pay insufficient attention to the opportunities to shape demand (see Section 6.4.4). For example, the marginal cost of reducing demand by 1 MW may be much less than supplying an additional 1 MW from even the cheapest resource. This can be addressed within the traditional problem framing by treating load reductions as a supply-side resource.

For each time interval (classically, each hour) the ensemble of generator contributions is to be determined by an algorithm that minimizes overall cost. The essential inputs are the marginal cost of each unit's output in dollars of fuel and operational expense per additional megawatt-hour, under the given circumstances. To a first approximation, megawatt-hours of electricity produced by various generators anywhere in the system are interchangeable. However, we know that based on their physical location within a transmission network, generators will differ in the amount of line losses incurred. A simple refinement accounts for the different impacts of supplying power from various locations by way of a *penalty factor* that effectively adjusts that generation unit's cost compared to others.

11.2.2 Lagrangian Method

The objective of economic dispatch is to determine a set of power contributions P_i from a set of m generators throughout the system—which, presumably, have different costs to operate—that collectively satisfies the system's power demand P_D at the lowest overall cost.

We begin by positing that each generator has a *cost curve* that reflects the hourly cost of operation, as a function of the power output level in megawatts. (In a competitive market environment, this would correspond to a generator's *bid*, stating that they are willing to supply some number of megawatts for a certain price.) Thus we define cost C_i in dollars per hour for generator i as a function of its output power P_i during that hour. The rate dC_i/dP_i at which the cost C_i increases with power output is called the *incremental cost*, sometimes abbreviated *IC*.

The generation cost may vary from one day or season to the next depending on circumstances, but here we assume it depends simply on that generator's current level of output during any given hour. In the simplest case, the cost curve might be a straight line, where the cost per megawatt-hour is independent of whether the unit is running at, say, 10%, 50%, or 90% of its rated capacity. In reality, the efficiency of a generator will vary over its range, and there is likely some fixed cost for running the generator at all.

To ensure a unique solution for economic dispatch, generator cost functions are required to be *convex*. In our case, where we are looking for a minimum, this means the cost curves are monotonically increasing, and with an increasing slope. It is not obvious why this should be physically realistic (e.g., why a generator should get less efficient as it approaches its maximum output), but keep in mind that this entire exercise is idealized.⁷

We can now write a total cost function C_T which is the sum total dollar costs of all m generators throughout the system, subject to the *constraint* that their total megawatt outputs must equal the total demand P_D , plus losses P_L . Mathematically, we state:

$$\begin{aligned} \min \quad & C_T = \sum_{i=1}^m C_i(P_i) \\ \text{s.t.} \quad & P_G = \sum_{i=1}^m P_i = P_D + P_L \end{aligned} \tag{11.2}$$

The general way to solve this type of problem is to take the derivative of the function to be minimized, with respect to each independent variable, and set it equal to zero. Such a derivative is called a *gradient*, denoted by the symbol ∇ (del or nabla), that includes partial derivatives with

⁷ To borrow a metaphor, we're saying that we may only drop our keys under the lamppost, so that we can search for them where the light is.

respect to each variable:

$$\nabla \mathbf{f}(\mathbf{x}) = \left[\frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_1}, \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_2}, \dots, \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_m} \right]$$

In our case, we want to take the partial derivative of C_T with respect to each of the P_i and set them all equal to zero. But how to account for the equality constraint? The method of Lagrangian multipliers restates that constraint in the form of a continuous variable that can be co-optimized along with the individual outputs, expressing the same problem as an *unconstrained optimization*. We define an abstract variable λ (lambda) called the *Lagrangian multiplier*, whose value is yet to be determined, but which should also be minimized. The total cost function C_T is augmented with a term involving λ to form the *Lagrangian*, written succinctly as

$$L(P_G, \lambda) = C_T + \lambda(P_D - P_G) \quad (11.3)$$

or more explicitly as

$$L(P_G, \lambda) = \sum_{i=1}^m C_i(P_i) + \lambda \left(P_D - \sum_{i=1}^m P_i \right) \quad (11.4)$$

A crucial argument can now be made: namely the very same set of P_i that minimize C_T will minimize L . Why? Because the λ term we have added to C_T to obtain L does not change the value of the function, as long as we are meeting the power balance constraint. Requiring that $P_G = P_D$ guarantees that $\lambda(P_D - P_G) = 0$, regardless of what value lambda might have. So under all relevant circumstances, we can take L as being synonymous with C_T , and the two functions ought to have the same minimum. This lets us take the partial derivative of L with respect to all the P_i , but also with respect to λ , and set them all equal to zero to find a necessary condition for the cost minimum.

Differentiating L with respect to each P_i introduces a λ term for each. Thus we obtain

$$\frac{\partial L(P_G, \lambda)}{\partial P_i} = \frac{dC_i(P_i)}{dP_i} - \lambda = 0 \quad \text{for all } i \quad (11.5)$$

This means that at the overall cost minimum, the incremental cost is exactly λ for every generator—consistent with the earlier statement that at the point where no further improvements are possible, the incremental costs should all be the same.

Moreover, we want this incremental cost λ to be as small as possible. Owing to the clever formulation of the Lagrangian, setting its partial derivative with respect to λ equal to zero reflects exactly our statement of power balance.

Having written out all the derivatives with respect to the various P_i (of which there are m), and the constraint that $P_D = P_G$, we are in a position to solve a set of $m + 1$ equations for $m + 1$ unknowns (the P_i and λ). If the generator cost functions are of polynomial form no higher than second order (i.e., containing exponents no greater than 2), we have a set of linear equations. Sets of linear equations with many variables are laborious but possible to solve by hand—even when there are many of them.

Two further considerations add a bit more complexity, but still leave the problem tractable. One is that each generator will be constrained by its rated capacity, or perhaps some other operational limit of available power. We don't know which constraints will come into play until we have solved the economic dispatch, and find that the desired MW output from a generator would exceed its limit. In that case, the generator output is pegged to its limit, and the unserved power has to be reallocated among the remaining generators.

Also, we may want to take into account that dispatching generators from different locations will have different effects on systemwide losses. For example, a generator with a low incremental cost may be a long distance away from major load centers, and getting more power from it would therefore raise overall system operating costs due to the increased transmission line losses (more so than procuring slightly more expensive power from nearby). But the systemwide losses aren't known without a candidate dispatch. The modern solution approach is *Optimal Power Flow* (Section 12.5.1), which cranks through the possibilities numerically.

The traditional, pencil-and-slide-rule modification for the closed-form Lagrangian method is to assign *penalty factors* to each generator. The penalty factor is based on the extent to which power dispatched from a given generator contributes to systemwide losses.⁸ Naturally, this will be an approximation that assumes some fixed system operating condition (i.e., power flow solution). The loss term P_L is in fact a function of the many P_i —except that one generator must serve as the *slack* (Section 12.2.7), so we omit P_1 and have only $m - 1$ independent variables. Augmenting Eq. (11.4) with the loss term, the Lagrangian cost function becomes

$$L(P_G, \lambda) = \sum_{i=1}^m C_i(P_i) + \lambda \left(P_D + P_L(P_2, P_3, \dots, P_m) - \sum_{i=1}^m P_i \right) \quad (11.6)$$

and its derivative with respect to the power P_i from the i th generator (starting with $i = 2$) becomes

$$\frac{\partial L}{\partial P_i} = \frac{dC_i}{dP_i} - \lambda \left(1 - \frac{\partial P_L}{\partial P_i} \right) = 0. \quad (11.7)$$

The penalty factor for the slack bus is just $L_1 = 1$. The penalty factor L_i for the i th generator is the factor which, when multiplied by its raw incremental cost, brings it to λ for achieving system optimality:

$$L_i \cdot \frac{dC_i}{dP_i} = \lambda \quad (11.8)$$

For example, a distant generator k whose recruitment incurs comparatively high losses would have a large $\partial P_L / \partial P_k$, and consequently a bigger penalty factor $L_k > 1$. For it to be successfully dispatched, its raw incremental cost dC_k / dP_k must be correspondingly less than λ . Combining Eqs. (11.7) and (11.8), we obtain

$$L_i = \frac{1}{1 - \frac{\partial P_L}{\partial P_i}} \quad (11.9)$$

The solution for a straightforward case without generator limits or losses is illustrated with an example below. Again, keep in mind that in reality, generator cost functions don't conform to these simplistic assumptions, nor any particular requirements such as convexity. In a competitive environment, generators make bids into a market, offering to sell some amount of power at some future time for some price—and their rationale for what to bid could be anything. The actual algorithms for clearing markets at minimum overall cost, which must consider multiple time horizons as well as physical network constraints, are much more complicated. But the exercise should help us appreciate how people managed power systems before the advent of computers.

⁸ Note how this framework is consistent with the vertically integrated business model, where system operating costs or savings are fungible across generation and transmission. The losses are simply added to the demand that needs to be met systemwide.

Example

Suppose that we need to supply 400 MW of load with three generators whose cost functions are given below. Find the least-cost dispatch values P_1^* , P_2^* and P_3^* , the incremental operating cost in \$/MWh, the total operating cost in \$/h, and the overall (average) generation cost in \$/MWh. Ignore losses and generator limits.

$$C_1 = 200 + 8P_1 + 0.01P_1^2 \text{ $/h}$$

$$C_2 = 250 + 8P_2 + 0.02P_2^2 \text{ $/h}$$

$$C_3 = 200 + 5P_3 + 0.05P_3^2 \text{ $/h}$$

The constrained optimization we are trying to solve is the following:

$$\min_{P_1, P_2, P_3} (C_1(P_1) + C_2(P_2) + C_3(P_3))$$

$$\text{subject to } P_1 + P_2 + P_3 = 400 \text{ MW}$$

The Lagrangian defining the unconstrained minimization is then:

$$L(P_1, P_2, P_3, \lambda) = (C_1(P_1) + C_2(P_2) + C_3(P_3)) + \lambda(400 - P_1 - P_2 - P_3)$$

Setting partial derivatives to zero gives us the following set of equations:

$$\frac{\partial L(P_1, P_2, P_3, \lambda)}{\partial P_1} = 8 + 0.02P_1^* - \lambda = 0$$

$$\frac{\partial L(P_1, P_2, P_3, \lambda)}{\partial P_2} = 8 + 0.04P_2^* - \lambda = 0$$

$$\frac{\partial L(P_1, P_2, P_3, \lambda)}{\partial P_3} = 5 + 0.1P_3^* - \lambda = 0$$

$$\frac{\partial L(P_1, P_2, P_3, \lambda)}{\partial \lambda} = 400 - P_1^* - P_2^* - P_3^* = 0$$

This system of four linear equations with four unknowns can be solved with a variety of methods. It is best to format into a matrix, which allows the information to be entered into standard solving tools:

$$\begin{bmatrix} 0.02 & 0 & 0 & -1 \\ 0 & 0.04 & 0 & -1 \\ 0 & 0 & 0.1 & -1 \\ -1 & -1 & -1 & 0 \end{bmatrix} \begin{bmatrix} P_1^* \\ P_2^* \\ P_3^* \\ \lambda^* \end{bmatrix} = \begin{bmatrix} -8 \\ -8 \\ -5 \\ -400 \end{bmatrix}$$

The solution for the set of optimal values P_i^* (with a bit of rounding error) is

$$P_1^* = 217.6 \text{ MW}$$

$$P_2^* = 108.8 \text{ MW}$$

$$P_3^* = 73.5 \text{ MW}$$

$$\lambda = 12.35 \text{ $/MWh}$$

at which point the incremental cost is the same throughout the system:

$$IC = \lambda = \frac{dC_1}{dP_1}(P_1^*) = \frac{dC_2}{dP_2}(P_2^*) = \frac{dC_3}{dP_3}(P_3^*) = 12.35 \text{ $/MWh}$$

The total hourly operating cost for the system is

$$C_1(P_1^*) + C_2(P_2^*) + C_3(P_3^*) = 2414 + 1357 + 838 = 4609 \text{ $/h}$$

The systemwide average generation cost on a per-megawatthour basis comes to

$$4609 \text{ \$/h} \div 400 \text{ MW} = 11.52 \text{ \$/MWh}$$

Because the incremental costs for all generators in this example are increasing, it makes sense that the average cost is less than the incremental cost of providing an additional MWh at the margin.

Problems and Questions

- 11.1** Discuss the information requirements for managing power demand and supply in a synchronous grid. Is it possible to balance supply and demand without knowing the total system load at any given moment? How could electric grids be operated over a century ago without telephone communications?
- 11.2** Consider the 1000-MW generator illustrated by the droop curve in Figure 11.3, initially operating at 600 MW and 60.00 Hz, with a regulation constant $R = 0.05 \text{ p.u.}$ Instead of load being added as in the example, suppose that 50 MW of load is suddenly lost.
- What is the frequency in Hz after the primary response?
 - Sketch the result of secondary frequency regulation on a graph, and indicate the desired final operating point.
- 11.3** An area has three generating units rated 200, 300, and 500 MVA, with regulation constants of 0.03, 0.04, and 0.05 p.u., respectively. Each unit is initially operating at one-half of its own rating.
- Find the frequency response characteristic β in MW/0.1 Hz, and expressed in per-unit on a 100-MVA base.
 - Find the change in system frequency and the change in mechanical power at each turbine, if the system load suddenly decreases by 150 MW.
 - Find the change in system frequency and the change in mechanical power at each turbine, if the system load suddenly increases by 100 MW.
 - List the assumptions necessary for solving this type of problem.
- 11.4** Consider Figure 11.9 showing the governor characteristics of two synchronous generators. Unit 1 is rated 1200 MW and Unit 2 is rated 600 MW. They are the only generators in the control area. The base frequency is 60 Hz.
- When the system is operating at 60.00 Hz, what are the respective outputs from Units 1 and 2, and what is the total load in MW?
 - What is the regulation constant R for each generator, in Hz/MW?
 - Within the operating range shown, what is the area frequency response characteristic β ?
 - If 300 MW of load (demand) is suddenly lost while the system is operating at 60.00 Hz, what is the new system frequency?
 - By how many MW does each unit reduce its output? Indicate the new operating points on the graph, approximately.
 - Is this problem describing primary, secondary or tertiary frequency regulation? How long would you expect this process to take, and does it require telecommunication between the generators?
 - Suppose that the control area is synchronously interconnected with another, larger area with many more generators. Qualitatively, how would this affect the problem?

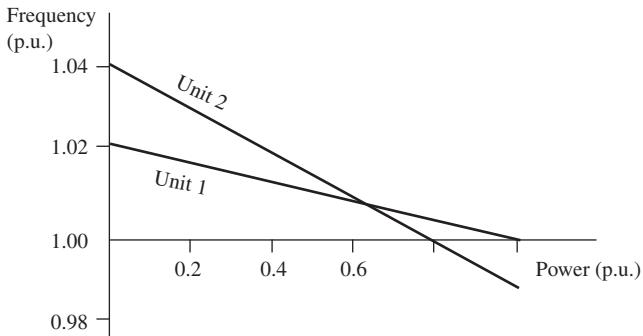


Figure 11.9 Droop curves for two generators.

- 11.5** A power system consists of two interconnected areas. Area 1 has 3000 MW of generation and an area frequency response characteristic $\beta_1 = 800 \text{ MW/Hz}$. Area 2 has 5000 MW of generation and an area frequency response characteristic $\beta_2 = 1600 \text{ MW/Hz}$. Each area is initially operating at one-half its rated generation, at 60.00 Hz and with zero tie line flow, when the load in Area 1 suddenly increases by 240 MW.

- (a) Find the frequency error and tie line flow error after primary frequency response in both areas.
- (b) Suppose that both areas perform secondary load frequency control (LFC). The frequency bias coefficient is set at $B_{f2} = \beta_2 = 1000 \text{ MW/Hz}$. What is the initial area control error (ACE) in each area, and what should happen subsequently to the system frequency, generation in each area, and the tie line flows?
- (c) Now suppose that Area 1 experienced not just a loss of load, but a communications blackout, and does not receive any ACE information. What happens to the system frequency, generation in each area, and the tie line flows in this scenario?
- (d) Comment on the advantage of interconnected areas in this context.

- 11.6** After World War II, East and West Germany were politically divided, including the city of Berlin itself. West Berlin was surrounded by the Berlin Wall. West Berlin's utility BEWAG operated its own 50-Hz electric grid as an island, geographically inside but electrically isolated from the German Democratic Republic (which was interconnected with Eastern Europe and the Soviet Union). Following the German reunification in 1990, BEWAG re-established a synchronous connection to the Western European UCTE grid in 1994.

- (a) Based on a population of about 2 million, roughly how large (order of magnitude) would you estimate the electric demand of West Berlin, in megawatts?
- (b) Figure 11.10a shows frequency versus time during normal grid operation, over a few minutes. (The vertical scale goes from 49.80 to 50.15 Hz.) Explain why it looks different before and after December 7. What changed as a result of the interconnection?
- (c) Figure 11.10b illustrates the behavior of grid frequency during three events in which a generator was lost (one of these occurred before and two occurred after 1994). Discuss the differences before and after the interconnection.
- (d) Based on Figure 11.10b, estimate the size of the UCTE system compared to West Berlin, by order of magnitude.

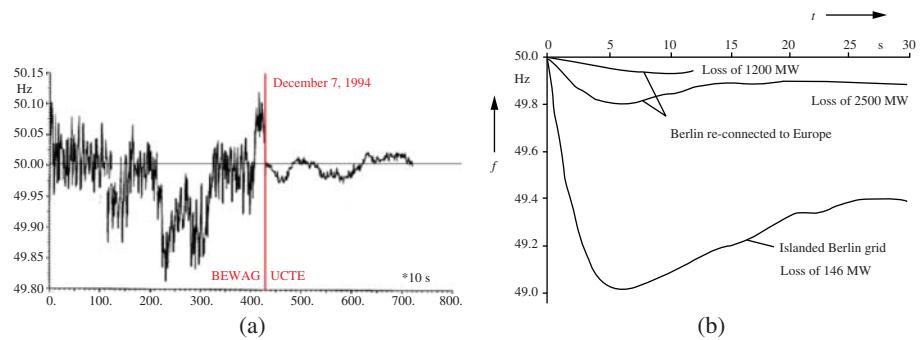


Figure 11.10 Frequency volatility (a) and nadir for loss-of-generation events (b) in West Berlin, before and after interconnection with Western Europe. Source: unknown.

11.7 Consider a system with two generators characterized by the following cost functions, where we will ignore generator limits:

$$C_1(P_1) = 400 + 5P_1 + 0.02P_1^2 \text{ \$/h}$$

$$C_2(P_2) = 800 + 5P_2 + 0.01P_2^2 \text{ \$/h}$$

- (a) Find the least-cost dispatch to meet a system demand of $P_D = P_1^* + P_2^* = 800 \text{ MW}$.
- (b) Find the incremental cost λ in \\$/MWh for each generator at this operating condition.
- (c) Find the average cost in \\$/MWh for energy provided by each generator. Do these values have to match? Explain.
- (d) Suppose the system load increases to 1000 MW. Can you tell by inspection which generator would be recruited to provide the additional 200 MW?
- (e) Suppose that Generator 2 in the previous problem is far from loads and entails a uniform 10% transmission losses, such that $P_L = 0.1P_2$. What is its penalty factor L_2 , and what is the least-cost dispatch for a system load of 1000 MW when taking losses into account?

12

Power Flow

12.1 Introduction

Power flow analysis is concerned with describing the operating state of an entire power system, by which we mean a network of generators, transmission lines, and loads that could represent an area as small as a municipality or as large as several states. Given certain known quantities—typically, the amount of power generated and consumed at different locations—power flow analysis allows one to determine other quantities. The most important of these quantities are the voltages at locations throughout the transmission system, which, for alternating current (a.c.), consist of both a magnitude and a time element or phase angle. Once the voltages are known, the currents flowing through every transmission link can be easily calculated. Thus, the name *power flow*, or *load flow*, as it is often called in the industry: given the amount of power delivered and where it comes from, power flow analysis tells us how it flows to its destination.

Owing to the peculiarities of a.c., but also to the sheer size and complexity of a real power system—its elaborate topology with many nodes and links, and the large number of generators and loads—it is no mean feat to deduce what is happening in one part of the system from what is happening elsewhere, despite the fact that these events are intimately related through well-understood, deterministic laws of physics. Although we can readily calculate voltages and currents for direct current (d.c.) circuits in terms of each other (as seen in Chapter 2), even a small network of a handful of a.c. power sources and loads defies our ability to write down easy formulas for the relationships among all the variables.

A key problem is that equations for power are *nonlinear*. Given voltages and currents, it is straightforward to find power, but given power generated and consumed at various locations, it is difficult to determine what the voltages and currents throughout a network must be in order to give rise to those power flows. Mathematicians would say that the system cannot be solved analytically; there is no *closed-form solution*. We can only get a numerical answer through a process of successive approximation or *iteration*. In order to determine its operating state, we must in effect simulate the entire system.

Historically, such simulations were accomplished through an actual miniature d.c. model of the power system in use. Generators were represented by small d.c. power supplies, loads by resistors, and transmission lines by appropriately sized wires. The voltages and currents could be found empirically by direct measurement. To find out how much the current on line A would increase, for example, due to Generator X taking over power production from Generator Y, one would simply adjust the values on X and Y and go read the ammeter on line A. The d.c. model does not exactly

match the behavior of the a.c. system, but it gives an approximation that is close enough for most practical purposes.

In the age of computers, we no longer need to physically build such models but can create them mathematically. With plenty of computational power, we can not only represent a d.c. system but also the a.c. system itself in a way that accounts for the subtleties of a.c. Such a simulation constitutes *power flow analysis*.

Power flow answers the question: What is the present operating state of the system, given certain known quantities? To do this, it uses a mathematical algorithm of successive approximation by iteration, or the repeated application of calculation steps. These steps represent a process of trial and error that starts with assuming one array of numbers for the entire system, comparing the relationships among the numbers to the laws of physics, and then repeatedly adjusting the numbers until the entire array is consistent with both physical law and the conditions stipulated by the user. In practice, this looks like a computer program to which the operator gives certain input information about the power system, and which then provides output that completes the picture of what is happening in the system—that is, how the power is flowing.

There are variations on what types of information are chosen as input and output, and there are also different computational techniques used by different programs to produce the output. Beyond the straightforward power flow program that simply calculates the variables pertaining to a single, existing system condition, there are more involved programs that analyze a multitude of hypothetical situations or system conditions and rank them according to some desired criteria; such programs are known as *optimal power flow* (OPF), discussed in Section 12.5.1.

This chapter is intended to provide the reader with a general sense of what power flow analysis is, how it is useful, and what it can and cannot do. Section 12.2 introduces the problem of power flow, showing how the power system is abstracted for the purpose of this analysis and how the known and unknown variables are defined. Section 12.3 discusses the interpretation of power flow results based on a sample case and points out some of the general features of power flow in large a.c. systems. Section 12.4 explicitly states the equations used in power flow analysis and outlines a basic mathematical algorithm used to solve the problem, including some simplifications or shortcuts. Section 12.5 addresses key applications of power flow analysis, and Section 12.6 considers the case of radial distribution systems.

12.2 The Power Flow Problem

12.2.1 Network Representation

In order to analyze any circuit, we use as a reference those points that are electrically distinct: that is, there is some impedance between them, which can sustain a potential difference. These reference points are called *nodes*. When representing a power system on a large scale, the nodes are called *buses*, since they represent an actual physical *busbar* where different components of the system are connected. A bus is electrically equivalent to a single point on a circuit, and it marks the location of one of two things: a generator that injects power, or a load that consumes power. At the degree of resolution generally desired on the larger scale of analysis, the load buses represent aggregations of loads (or very large individual industrial loads) at the location where they connect to the high-voltage transmission system. Such an aggregation may in reality be a transformer connection to a subtransmission system, which in turn branches out to a number of distribution substations; or it may be a single distribution substation from which originate a set of distribution

feeders (see Figure 7.4).¹ In any case, whatever lies behind the bus is taken as a single load for purposes of the power flow analysis.

The buses in the system are connected by transmission lines. At this scale, one does not generally distinguish among the three phases of an a.c. transmission line (Section 4.1). Based on the assumption that, to a good approximation, the same thing is happening on each phase, the three are condensed by the model into a single line, making a so-called *one-line diagram*. Indeed, a single line between two buses in the model may represent more than one three-phase circuit. Still, for this analysis, all the important characteristics of these conductors can be condensed into a single quantity, the *impedance* of the one line (see Section 3.3). Since the impedance is essentially determined by the physical characteristics of the conductors (such as their material composition, diameter, and length), it is taken to be constant.² Note that this obviates the need for geographical accuracy, since the distance between buses is already accounted for within the line impedance, and the lines are drawn in whatever way they best fit on the page or the screen.

Thus, the model so far represents the existing hardware of the power system, drawn as a network of buses connected by single lines. An example of such a one-line diagram is shown in Figure 12.1, to illustrate the level of abstraction. This topology or characteristic connection of the network can be changed by switching operations, whereby, for example, an individual transmission line can be taken out of service. Such changes, of course, must be reflected by redrawing the one-line diagram, where now some lines may be omitted or assigned a new impedance value. For a given analysis run, though, the network topology is taken to be fixed.

The mathematical representation of a transmission network in the *bus admittance matrix* is found in Section 12.4.2.

12.2.2 Choice of Variables

From the analysis of simple d.c. circuits in Chapter 2, we are familiar with the notion of organizing the descriptive variables of the circuit into categories of “knowns” and “unknowns,” whose relationships can subsequently be expressed in terms of multiple equations. Given sufficient information, these equations can then be manipulated with various techniques so as to yield numerical results for the hitherto unknowns. As readers may recall from high school algebra, the conditions under which such a system of equations is solvable (meaning that it can yield unambiguous numerical answers) are straightforward: there must be exactly one equation for each unknown quantity. Each equation represents one statement relating one unknown variable to some set of known quantities. This set of equations must not be redundant: if any one equation duplicates information implied by the others, it does not tell us anything new and therefore does not count toward making the whole system solvable. If there are fewer equations than unknowns, we do not have enough information to decide which values the unknowns must take (in other words, the information given does not rule out a multiplicity of possibilities); if there are more equations than unknowns, the system is overspecified, meaning that some equations are either redundant or

1 Of course, it is possible to run power flow analysis at different scales, including a smaller scale that explicitly incorporates more distribution system elements. In the present discussion, we emphasize the largest transmission scale because of its general and economic importance. Also note that distribution systems usually submit to simpler methods of analysis because of their radial structure, implying that power flows in only one direction. Power flow analysis is indispensable, however, for the meshed networks that characterize the transmission system.

2 Ambient conditions such as conductor temperature are hard to know exactly, but have a small enough effect on line impedance that they are usually neglected.

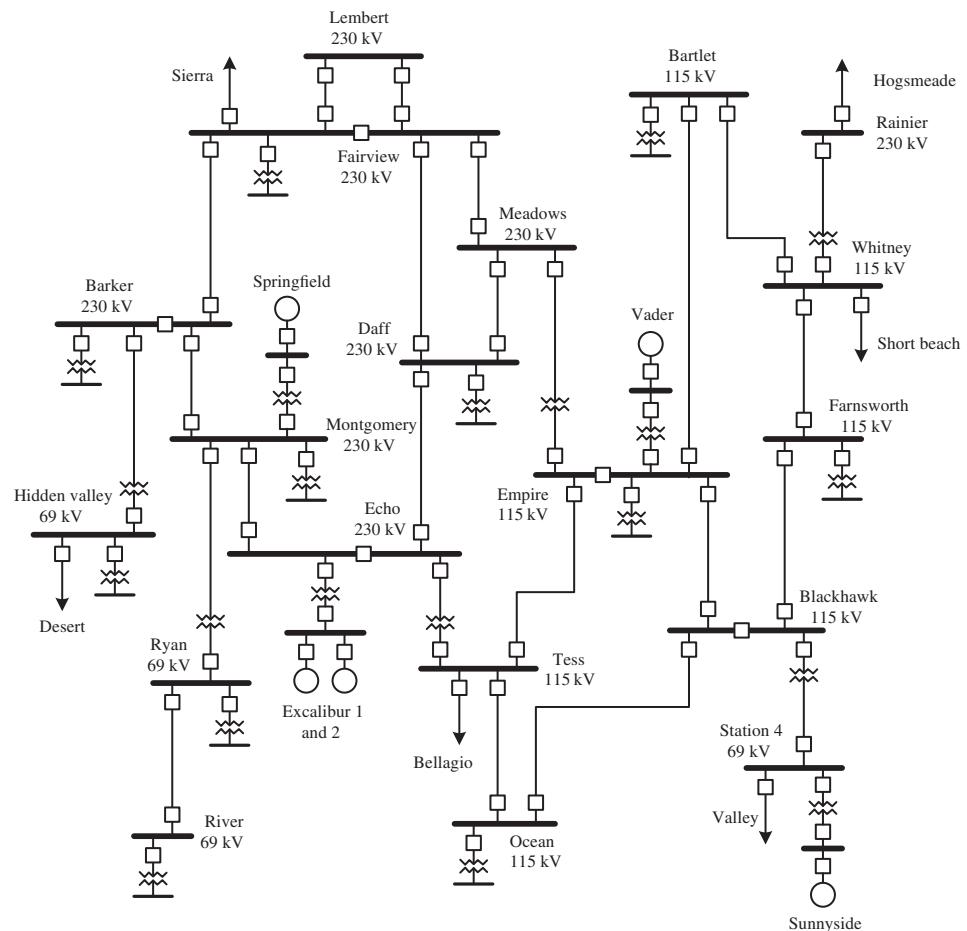


Figure 12.1 One-line diagram for a fictitious power system.

mutually contradictory. In order to determine whether an unambiguous, unique solution to a system of equations such as those describing an electric power system can be found, one must begin by taking an inventory of variables and information that translates into equations for those variables.

As introduced in Chapter 1, there are two basic quantities that describe the flow of electricity: voltage and current. Recognizing these quantities in simple d.c. circuits in Chapter 2, we saw that both voltage and current will vary from one location to another in a circuit, but they are everywhere related: the current through each circuit branch corresponds to the voltage or potential difference between the two nodes at either end, divided by the impedance of this branch. It is generally assumed that the impedances throughout the circuit are known, since these are more or less permanent properties of the hardware. Thus, if we are told the voltages at every node in the circuit, we can deduce from them the currents flowing through all the branches, and everything that is happening in the circuit is completely described. If one or more pieces of voltage information were missing, but we were given appropriate information about the current instead, we could still work backwards and solve the problem. In this sense, the number of variables in a circuit corresponds to the number of electrically distinct points in it: assuming we already know all the properties of the

hardware, we need to be told one piece of information per node in order to figure out everything that's going on in a d.c. circuit.

For a.c. circuits, the situation is more complicated because we have introduced the dimension of time to capture an ongoing oscillation or movement. Thus each of the two main variables, voltage and current, has two numerical components: a magnitude component and a time component. To fully describe the voltage at any given node in an a.c. circuit, we must therefore specify two numbers: a voltage magnitude and a voltage angle (Section 3.1). Given the complex impedances (Section 3.3) of all the network branches, which are also number pairs, we can solve for the current magnitude and angle in each branch. The product of voltage and current gives the amount of power transferred at any point, which is again a pair of numbers with a real and a reactive component (Section 3.4.2). Thus, an a.c. circuit requires exactly two pieces of information per node for its operating state to be completely determined. More than two, and they are either redundant or contradictory; fewer than two, and possibilities are left open so that the system cannot be solved.

A word of caution is in order: Owing to the *nonlinear* nature of the power flow problem, it may be impossible to find one unique solution even with the proper number of equations, as more than one answer fits the given configuration. Nonlinearity is further discussed in Section 12.2.3. In most situations, it is straightforward to identify the “correct” solution among the mathematical possibilities based on physical plausibility and common sense. Conversely, there may be no solution at all, if the given information does not correspond to an actual physical situation.

Having discussed voltage and current, each with magnitude and angle, as the basic electrical quantities, which are known and which are unknown? In practice, current is not known; the currents through the various circuit branches turn out to be the last thing that we calculate once we have completed the power flow analysis. Voltage, as we will see, is known explicitly for some buses but not for others. More typically, what is known is the amount of *power* going into or out of a bus. Power flow analysis takes all the known real and reactive power flows at each bus, and those voltage magnitudes that are explicitly known, and from this information calculates the remaining voltage magnitudes and all the voltage angles. This is the hard part. The easy part, finally, is to calculate the current magnitudes and angles from the voltages.

From Section 3.4, we know how to calculate real and reactive power from voltage and current: power is basically the product of voltage and current, and the relative phase angle between voltage and current determines the respective contributions of real and reactive power. Conversely, one can deduce voltage or current magnitude and angle if real and reactive power are given, but it is far more difficult to work out mathematically in this direction. This is because each value of real and reactive power would be consistent with many different possible combinations of voltages and currents. In order to choose the correct ones, we have to check each node in relation to its neighboring nodes in the circuit and find a set of voltages and currents that are consistent all the way around the system. This is what power flow analysis does.

12.2.3 Nonlinearity

A linear function is one in which the output is proportional to the inputs. One consequence is that the principle of *superposition* holds: when more input is added, the new output is the sum of the outputs due to the original plus the newly added input. That is, for the linear function $f(x) = c x$ it is always true that

$$f(x_1 + x_2) = c(x_1 + x_2) = c x_1 + c x_2 = f(x_1) + f(x_2)$$

The same is not true for the nonlinear function $g(x) = c x^2$ where

$$g(x_1 + x_2) = c (x_1 + x_2)^2 \neq c x_1^2 + c x_2^2$$

We have already encountered superposition in Section 2.4, where we describe linear aspects of circuit behavior—specifically, Ohm's law. For linear circuit elements (including resistors, inductors, and capacitors), voltage is a linear function of current ($V = IZ$) and vice versa ($I = YV$), with the impedance Z or admittance Y playing the role of the proportionality constant. As a result, the voltage resulting from an added current source in a circuit can be added to the original voltage; or the current due to an added voltage source can be added to the original current.

For example, in Figure 12.2 (very similar to Figure 2.6), the current in the center branch of the circuit is the sum of currents due to the two voltage sources on either side. The reason we may simply add currents is because the behavior of the center branch is characterized by a linear equation: $V = IR$, where the resistance is a known constant.

Now suppose the center branch is characterized not by a constant impedance, but by a certain amount of power demand (i.e., it is modeled as a *constant power load*). Ohm's law still applies, but the impedance in it is no longer some known constant value. Instead, the impedance may shift around depending on the voltage conditions, so as to keep the product of voltage and current for that load constant. We can write an equation for power, $S = I^*V$, which at first glance might appear linear.³ However, this equation is not in fact linear because V itself is a function of I , or I of V . If we substitute Ohm's law $V = IZ$ into the equation for power, we get

$$S = I^*IZ \quad \text{or} \quad S = \frac{VV^*}{Z^*}$$

which is quadratic in either I or V . Consequently, it is not at all obvious how to solve for the voltage across or current through the center circuit branch in Figure 12.3.

Kirchhoff's laws still apply, but don't entirely solve our problem. Although we can still write $I_3 = I_1 + I_2$, the actual values of I_1 and I_2 cannot be determined independently of the voltage V_3 at the central circuit node. Likewise, V_3 can be written as the difference between the voltage source and the voltage drop across the resistance on either side of the circuit, but those voltage drops cannot be determined independently of the current.

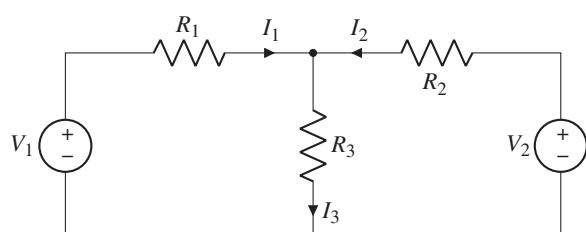


Figure 12.2 Current I_3 is found by adding currents $I_1 + I_2$, superimposing the two partial circuits on the center branch. This works because all circuit elements have linear characterizations.

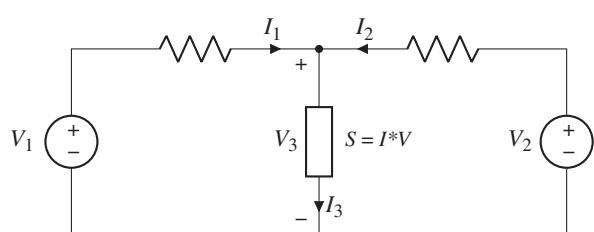


Figure 12.3 In a nonlinear system, we cannot directly solve for V_3 or I_3 .

³ The asterisk denotes the complex conjugate; see Section 3.5.7.

What to do? We could make a guess at the current I_3 , see what voltage V_3 it implies, and then check how consistent this is with V_1 and V_2 . Or we could guess at V_3 and check the corresponding currents; if they don't add up, we can adjust our guess for V_3 . With repeated guesses, we can try to get the discrepancy to shrink until the solution is both internally consistent and matches all given information. Such a brute-force, trial-and-error approach may seem desperate, but it is in fact how the power flow problem is handled professionally—by the fancier name of *iterative solution methods*.

12.2.4 Types of Buses

Let us now articulate which variables will be given for each bus as inputs to the analysis. Here we must distinguish between different types of buses based on their actual, practical operating constraints. The two main types are *generator buses* and *load buses*. At a load bus, we assume that the power consumption is given—determined by the consumer—and we specify two numbers, real and reactive power, for each load bus.⁴ Referring to the symbols P and Q for real and reactive power, load buses are referred to as P,Q buses in power flow analysis.

At the generator buses we could in principle also specify P and Q . Here we run into two issues: one has to do with balancing the real power needs of the system, and the other with managing reactive power. To address real power balance, we will specify P for all but one generator, the slack bus. To represent the actual operational control for most generators, we typically use the generator bus voltage V instead of the reactive power Q as the second variable. Most generator buses are therefore called P,V buses.

12.2.5 Variables for Balancing Real Power

Balancing the system means that all the generators in the system collectively must supply power in exactly the amount demanded by the load, plus the amount lost on transmission lines. This applies to both real and reactive power, but let us consider only real power first. If we tried to specify a system in which the sum of P generated did not match the P consumed, our analysis would yield no solution, reflecting the fact that in real life the system would lose synchronicity and crash. Therefore, for all situations corresponding to a stable operation of the system, and thus a viable solution of the power flow problem, we must require that real power generated and consumed matches up. Of course, we can vary the contributions from individual generators—that is, we can choose a different *dispatch*—so long as the *sum* of their P 's matches the amount demanded by the system. As mentioned earlier, this total P must not only match the load demand, it must actually *exceed* that amount in order to make up for the transmission losses, which are the resistive I^2R energy losses (Section 1.4).

Now we have a problem: How are we supposed to know ahead of time what the transmission losses are going to be? Once we have completed the power flow analysis, we will know what the current flows through all the transmission lines are going to be, and combining this information

⁴ We assume that the load's power demand is independent of the voltage at that bus. This may seem to fly in the face of everything said in Section 6.3.1. However, we are talking here about the voltage magnitude at a bus as modeled in the transmission system, which would typically represent a substation, which is not the actual service voltage where customer loads are connected. Owing to voltage regulating equipment in the distribution system that aims to keep the service voltage constant (Section 7.4), it is fair to assume that the actual service voltage to customers is independent of the transmission bus voltage, and that the power drawn at the bus should remain independent of bus voltage magnitude.

with the known line impedances will give us the losses. But we cannot tell *a priori* the amount of losses. The exact amount will vary depending on the dispatch, or amount of power coming from each generator, because a different dispatch will result in a different distribution of current over the various transmission paths, and not all transmission lines are the same. Therefore, if we were given a total P demanded at the load buses and attempted now to set the correct sum of P for all the generators, we could not do it.

The way to deal with this situation mathematically resembles the way it would be handled in actual operation. Knowing the total P demanded by the load, we could begin by assuming a typical percentage of losses (say, 5%). We now dispatch all the generators in the system so that the sum of their output approximately matches what we expect the total real power demand (load plus losses) to be (in this case, 105% of load demand). But since we do not yet know the exact value of the line losses for this particular dispatch (seeing that we have barely begun our power flow calculation), we will probably be off by a small amount. A different dispatch might, for example, result in 4.7% or 5.3% instead of 5% losses overall. We now make the plausible assumption that this uncertainty in the losses constitutes a sufficiently small amount of power that a single generator could readily provide it. So we choose one generator whose output we allow to adjust, depending on the system's needs: we allow it to "take up the slack" and generate more power if system losses are greater than expected, or less if they are smaller. In power flow analysis, this one generator bus is appropriately labeled the *slack bus*, or sometimes *swing bus*.

Thus, as the input information to our power flow analysis, we specify P for *one less* than the total number of buses. What takes the place of this piece of information for the last bus is the requirement that the system remain balanced. This requirement will be built into the equations used to solve the power flow and will ultimately determine what the as yet unknown P of the slack bus has got to be. The blank space in the initial specifications for the slack bus, where P is not given, will be filled by the voltage angle, to be discussed in Section 12.2.7.

12.2.6 Variables for Balancing Reactive Power

Analogous to real power, the total amount of reactive power generated throughout the system must match the amount of reactive power consumed by the loads.⁵ Whereas in the case of a mismatch of real power, the system loses synchronicity, a mismatch of reactive power would lead to voltage collapse. Also analogous to real power transmission losses, there are *reactive power losses*. Reactive losses are defined simply as the difference between reactive power generated and reactive power consumed by the metered load.

Physically, these losses in Q reflect the fact that transmission lines have some reactance (Section 3.3) and thus tend to "consume" reactive power; in analogy to I^2R , we could call them I^2X losses. The term "consumption," however, like the reactive power "consumption" by a load, does not directly imply an energy consumption in the sense of energy being withdrawn from the system. To be precise, the presence of reactive power does necessitate the shuttling around of additional current, which in turn is associated with some real I^2R losses "in transit" of a much smaller magnitude. But these second-order I^2R losses (the side effect of a side effect) are already

⁵ Recall that in Section 3.4, we put quotation marks around the terms "supplied" and "consumed" as they apply to reactive power, since this is a somewhat arbitrary nomenclature. No net energy is produced or consumed by either generator or load due to reactive power exchange. However, instantaneous power must be balanced throughout the system at all times during each cycle. This is accomplished by balancing both P and Q . Physically, if instantaneous power is imbalanced, the difference is made up from the potential energy stored in electric and magnetic fields throughout the system. If this finite capacity were exhausted, voltage would collapse.

captured in the analysis of real power for the system. The term “reactive losses” thus does not refer to any physical measure of something lost, but rather should be thought of as an accounting device. While real power losses represent physical heat lost to the environment and therefore always have to be positive,⁶ reactive losses on a given transmission link can be positive or negative, depending on whether inductive or capacitive reactance plays a dominant role.

In any case, what matters for both operation and power flow analysis is that Q , just like P , needs to be balanced at all times. Thus, just as for real power, all the generators in the system must generate enough reactive power to satisfy the load demand *plus* the amount that vanishes into the transmission lines.

This leaves us with the analogous problem of figuring out how much total Q our generators should produce, not knowing ahead of time what the total reactive losses for the system will turn out to be: as with real losses, the exact amount of reactive losses will depend on the dispatch. Operationally, though, the problem of balancing reactive power is considered in very different terms. When an individual generator is instructed to provide its share of reactive power, the control objective is usually expressed in terms of maintaining a certain voltage magnitude at the generator bus, rather than injecting a certain number of MVAR. An automatic voltage regulator (AVR) continually and automatically adjusts through the generator’s field current (Section 10.4.2) and thereby alters the reactive power output.

Generator bus voltage magnitude is a relatively straightforward variable to control. It is also an indicator of whether the correct amount of reactive power is being generated throughout the system. When the combined generation of reactive power by all the generators matches the amount consumed, their bus voltages hold steady. Conversely, if there is a need to increase or decrease reactive power generation, adjusting the field current at one or more generators so as to return to the voltage set point will automatically accomplish this objective.⁷ The new value of MVAR produced by each generator can then be read off the dial for accounting purposes.

Conveniently for power flow analysis, then, there is no need to know explicitly the total amount of Q required for the system. Specifying the voltage magnitude is essentially equivalent to requiring a balanced Q . In principle, we could specify P and Q for each generator bus, except for one slack bus assigned the voltage regulation (and thus the onus of taking up the slack of reactive power). For this “reactive slack” bus we would need to specify voltage magnitude V instead of Q , with the understanding that this generator would adjust its Q output as necessary to accommodate variations in reactive line losses. In practice, however, since voltage is already the explicit operational control variable, it is customary to specify V instead of Q for all generator buses, which are therefore called P, V buses.⁸ In a sense, this assignment implies that all generators share the “reactive slack,” in contrast to the real slack that is taken up by only a single generator.

12.2.7 The Slack Bus

We have now, for our power flow analysis, three categories of buses: P, Q buses, which are generally load buses, but could in principle also be generator buses; P, V buses, which are necessarily generator buses (since loads have no means of voltage control); and then there is the slack bus, for which we cannot specify P , only V . What takes the place of P for the slack bus?

⁶ Lest we violate the second law of thermodynamics, which forbids heat from flowing spontaneously from the air into the wire and making electricity.

⁷ See Section 10.4.4 for more about how generators share MVAR load.

⁸ It is important to remember that the V of the P, V bus represents only voltage magnitude, not angle. To avoid any confusion, the careful notation $P, |V|$ is sometimes used, where the vertical lines indicate magnitude.

As introduced in Section 10.4.1, *real* power balance manifests operationally as a steady frequency such as 60 Hz. A constant frequency is indicated by an unchanging voltage angle, which for this reason is also known as the *power angle*, at each generator. When more power is consumed than generated, the generators' rotation slows down: their electrical frequency drops, and their voltage angles fall farther and farther behind. Conversely, if excess power is generated, frequency increases and the voltage angles move forward. While generators are explicitly dispatched to produce a certain number of megawatts, the necessary small adjustments to balance real power in real time are made (by at least one or more *load-following* generators) through holding the generator frequency steady at a specified value. Not allowing the frequency to depart from this reference value is equivalent to not letting the voltage angle (relative to the rotating reference frame) to increase or decrease over time.

In power flow analysis, the slack bus is the one mathematically assigned to do the load following. Its instructions, as it were, are to do whatever is necessary to maintain real power balance in the system. Physically, this would mean holding the voltage angle constant. The place of P will therefore be taken by the *voltage angle*, which is the variable that in effect represents real power balance. We can think of the voltage angle here as analogous to the voltage magnitude in the context of reactive power. Specifying that bus voltage magnitude should be kept constant effectively amounts to saying that whatever is necessary should be done to keep the system reactive power balanced. Likewise, specifying a constant voltage angle at the generator bus amounts to saying that this generator should do whatever it takes to keep real power balanced.

We thus assign to the slack bus a voltage angle, which, in keeping with the conventional notation for the context of power flow analysis, we will call θ (lowercase Greek theta). This θ can be interpreted as the relative position of the slack bus voltage at time zero. It is the same quantity that is elsewhere called the *power angle* and labeled as δ .

What is important to understand here is that the actual numerical value of this individual voltage phase angle has physical meaning only in relation to a reference. It is the *difference* between the voltage angle at one bus and another, as well as its rate of change, that matters. Physically, the angle difference between voltage curves at two locations corresponds to the difference in the precise timing of the zero crossings (or voltage maxima), which in turn is related to the power transfer between those two locations in the network.⁹ Also, a fixed value of θ has physical meaning in that it implies this angle will not change as the system operates. The choice of a particular numerical value for the first θ in a network amounts to a choice of coordinate system (i.e., what do we call Time Zero). Given this reference, the voltage angles for each of the other buses throughout the system will take on different values depending on their contribution to real power. But as long as the entire system is in a state of equilibrium, where generation equals load, these angles will hold steady.

We now conveniently take advantage of the slack bus to establish a systemwide reference for timing, and we might as well make things simple and call the reference point "zero." This could be interpreted to mean that the alternating voltage at the slack bus has its maximum at the precise instant that we depress the "start" button of an imaginary stopwatch, which starts counting the milliseconds (in units of degrees within a complete cycle of 1/60th second) from time zero. In principle, we could pick any number between 0° and 360° as the voltage angle for the slack bus, but 0° is the simple and conventional choice.

⁹ Since the voltage continually alternates, it would be of little use to say that the voltage maximum at Bus A occurs at exactly 3:00:00 P.M. Rather, we would want to know that the voltage maximum at Bus A always occurs one-tenth of a cycle later than at Bus B.

Table 12.1 Variables in power flow analysis.

Type of Bus	Variables Given (Knowns)	Variables Found (Unknowns)
Generator	Real power (P)	Voltage angle (θ)
	Voltage magnitude (V)	Reactive power (Q)
Load or generator	Real power (P)	Voltage angle (θ)
	Reactive power (Q)	Voltage magnitude (V)
Slack	Voltage angle (θ)	Real power (P)
	Voltage magnitude (V)	Reactive power (Q)

12.2.8 Summary of Variables

To summarize, our three types of buses in power flow analysis are P, Q (load bus), P, V (generator bus), and θ, V (slack bus). Given these two input variables per bus, and knowing all the fixed properties of the system (i.e., the impedances of all the transmission links, as well as the a.c. frequency), we now have all the information required to completely and unambiguously determine the operating state of the system. This means that we can find values for all the variables that were not originally specified for each bus: θ and V for all the P, Q buses; θ and Q for the P, V buses; and P and Q for the slack bus. The known and unknown variables for each type of bus are listed in Table 12.1 for easy reference.

Once we know θ and V , the voltage angle and magnitude, at every bus, we can very easily find the current through every transmission link; it becomes a simple matter of applying Ohm's law to each individual link. (In fact, these currents have already been determined implicitly, so that by the time the program announces θ 's and V 's, all the hard work is done.) Depending on how the output of a power flow program is formatted, it may state only the basic output variables, as in Table 12.1, it may explicitly state the currents for all transmission links in amperes; or it may express the flow on each transmission link in terms of an amount of real and reactive power flowing, in megawatts (MW) and MVAR.

12.3 Example with Interpretation of Results

12.3.1 Six-bus Example

Consider the six-bus example illustrated in Figure 12.4.¹⁰

This example is simple enough for us to observe in detail, yet too complex to predict its behavior without numerical power flow analysis.

Each of the six buses has a load, and four of the buses also have generators. Bus 1, keeping with convention, is the slack bus. Buses 2–4, which have both generation and loads, are modeled as P, V buses; the local load is simply subtracted from the real and reactive generation at each. Buses 5 and 6, which have only loads, are modeled as P, Q buses.

¹⁰ This example is taken from PowerWorld™, a power flow software application available from www.powerworld.com for both commercial and educational use; a short version is available as a free download. The case illustrated here is from the menu of the standard demonstration cases in PowerWorld, labeled "Contour 6-Bus" (retrieved October 2004).

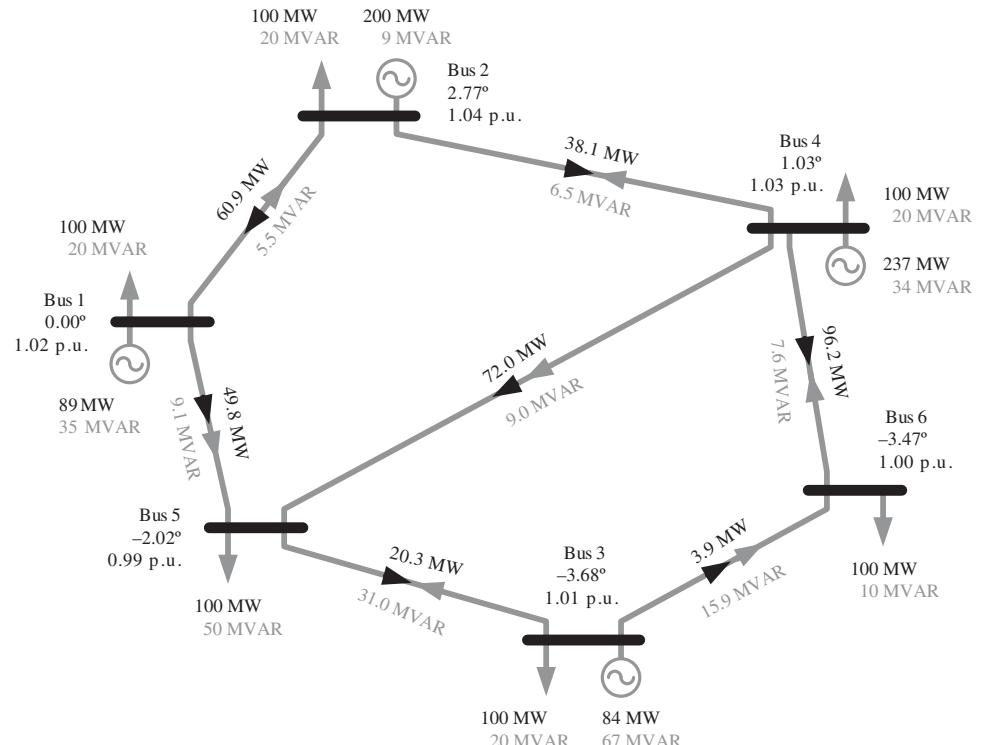


Figure 12.4 Six-bus power flow example.

The distribution of loads and the generation dispatch, for both real and reactive power, are completely determined somewhere outside the power flow program, whether in the real world or the program user's fantasy. The one exception is the generator at the slack bus, whose real power output varies so as to accommodate systemwide losses. In addition to the MW and MVAR loads and the MW generation levels for every generator (except the slack), the user specifies the voltage magnitudes to be maintained at each generator bus; the program then computes the MVAR generation necessary to maintain this voltage at each bus. (It is also possible to specify MVAR generation and allow the program to determine the voltage magnitudes, but, as mentioned earlier, the former method better resembles real-life operations.)

By convention, the voltage angle at the slack bus is set to 0.00° . The power flow program computes the voltage angle at each of the other five buses in relation to the slack bus. We may now begin to observe the relationship between real power and voltage angle: a more positive voltage angle generally corresponds to an injection of power into the system and a more negative voltage angle to a consumption of real power. Buses 2 and 4, which both have generation exceeding local load, have positive voltage angles of 2.77° and 1.03° , respectively. Bus 3, though it has a generator, is still a net consumer of real power, with 100 MW load and only 84 MW generated; its voltage angle is -3.68° . Buses 5 and 6 have loads only and voltage angles of -2.02° and -3.47° , respectively.

Note, however, that the voltage angles are not in hierarchical order depending on the amount of power injected or withdrawn at each individual bus. This is because we also must consider the location of each bus relative to the others in the system and the direction of power flow between them. For example, consider Buses 2 and 4. Net generation at Bus 4 is greater than at Bus 2 (137 MW compared to 100 MW), yet the voltage angle at Bus 2 is more positive. We can see that this is due to

the location of these buses in the system: real power is generally flowing from north to south, that is, from Bus 2 to the neighborhood of Buses 5, 3, and 6 where there is more load and less generation. As indicated by the black arrow on the transmission link, real power is flowing from Bus 2 to Bus 4. As a rule, real power flows from a greater to a smaller voltage angle. This rule holds true for six of the seven links in this sample case; the exception is Link 3–6, where both the power flow and the difference in voltage angle are very small. The reader can verify that throughout this case, while power flow and voltage angle are not exactly proportional, a greater flow along a transmission link is associated with a greater angle difference.

We now turn to the relationship between reactive power and voltage magnitude, which is similar to that between real power and voltage angle. The nominal voltage of this hypothetical transmission system is 138 kV. However, just as the timing or angle of the voltage differs by a small fraction of a cycle at different locations in the grid, the magnitude, too, has a profile across the system with different areas a few percent higher or lower than the nominal value. Because it is this percentage, not the absolute value in volts, that is most telling about the relationship among different places in the grid, it is conventional to express voltage magnitude in *per-unit* terms (see Section 8.7). Per-unit (p.u.) notation simply indicates the local value as a multiple of the nominal value; in this case, 138 kV equals 1.00 p.u. The voltage magnitude at Bus 1 is given as 1.02 p.u., which translates into 141 kV; at Bus 5, the voltage magnitude of 0.99 p.u. means 137 kV.

As a rule, reactive power tends to flow in the direction from greater to smaller voltage magnitude. In our example, this rule holds true only for the larger flows of MVAR, along Links 1–5, 3–5, 4–5, and 3–6. The reactive power flows along Links 1–2, 4–2, and 6–4 do not follow the rule, but they are comparatively small.

Note that real and reactive power do not necessarily flow in the same direction on a given link. This should not be surprising, because the “direction” of reactive power flow is based on an arbitrary definition of the generation or consumption of VARs; there is in fact no net transfer of energy in the direction of the gray arrow for Q . Also, note that having Q flow opposite P does not imply any “relief” or reduction in current. For example, on Link 3–5, the real power flow P is 20.3 MW and reactive flow Q is 31.0 MVAR. In combination, this gives apparent power S of 37.1 MVA (using $S^2 = P^2 + Q^2$), regardless of the direction of Q . (Recall that MVA are the relevant units for thermal line loading limits, since total current depends on apparent power.)

From Figure 12.4, it is possible to evaluate the total real and reactive system losses, simply by observing the difference between total generation and total load. The four generators are supplying 89, 200, 84, and 237 MW, respectively, for a total of 610 MW of real power generated. Subtracting the six loads of 100 MW each, the total real power losses throughout the transmission system for this particular scenario are therefore 10 MW. On the reactive side, total generation is 145 MVAR, while total reactive load is 140 MVAR, and system reactive losses amount to roughly 5 MVAR.

The discerning reader may have noticed, however, that the stated line flows in Figure 12.4, which are average values for each link, cannot all be reconciled with the power balance at each bus. To account for losses in a consistent fashion, we must record both the power (real or reactive) entering and exiting each link. In Figure 12.5, these data are given for real power (MW) in black and reactive power (MVAR) in gray. The numbers in parentheses represent the losses, which are the difference between power flows at either end. Bus power, line flows, and losses are rounded to different decimal places, but the numbers do add up correctly for each bus and each link.

The most significant losses tend to occur on links with the greatest power flow. In this case, Link 4–6 has the greatest power flow with 96.2 MW real and 7.6 MVAR reactive, yielding 96.5 MVA apparent, and the greatest losses. While the real line losses are all positive, as they should be, the negative signs on some of the reactive losses indicate negative losses; we might consider them

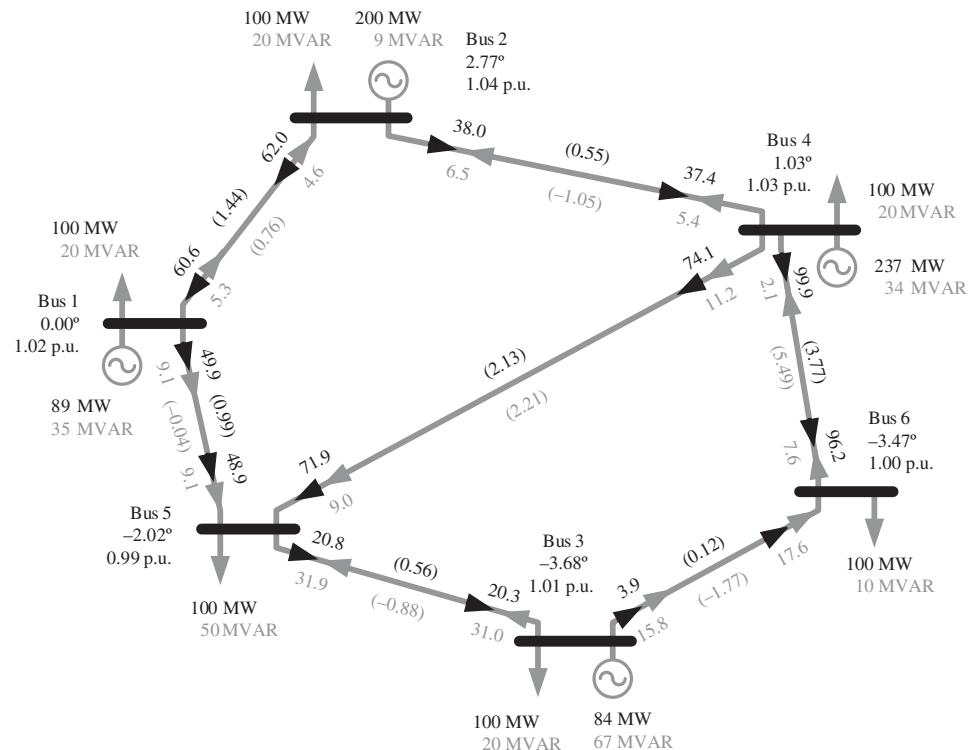


Figure 12.5 Six-bus power flow example with losses.

“gains,” although nothing is actually gained. Reactive losses depend on operating conditions and impedance, where the model of a transmission link may incorporate reactive compensation such as capacitors. It is typical for system reactive losses to be positive overall, as they are in this example. Like real losses, reactive losses are related to the current and therefore apparent power flow. Thus, we also observe the greatest reactive losses in our example on Link 4–6. The real and reactive losses for every link can be totaled to confirm the estimated system losses given earlier.

12.3.2 Tweaking the Case

To gain a better sense of a power system’s behavior and the information provided by power flow modeling, let us now make a small change to the operating state in the six-bus example and observe how the model responds. We simply increase the load at Bus 5 by 20% while maintaining the same power factor, thus changing it from 100 MW real and 50 MVAR reactive to 120 MW real and 60 MVAR reactive. This change is small enough for the generator at the slack bus to absorb, so we need not specify increased generation elsewhere. Indeed, generation at Bus 1 increases from 89 to 110 MW. Note that the difference amounts to 21, not 20 MW, as the increased load also entails some additional losses in the system. The new scenario is illustrated in Figure 12.6.

As we would expect, the line flows to Bus 5 increase by a total of 20 MW. The bulk (about 14.5 MW) of this additional power comes from Bus 1, about 4 MW from Bus 4, and the balance appears as a reduction of about 1.5 MW in the flow to Bus 3.

The changes do not stop here, however; they have repercussions for the remainder of the system. Three of the other buses are defined as P, V buses, and therefore have fixed voltage magnitudes.

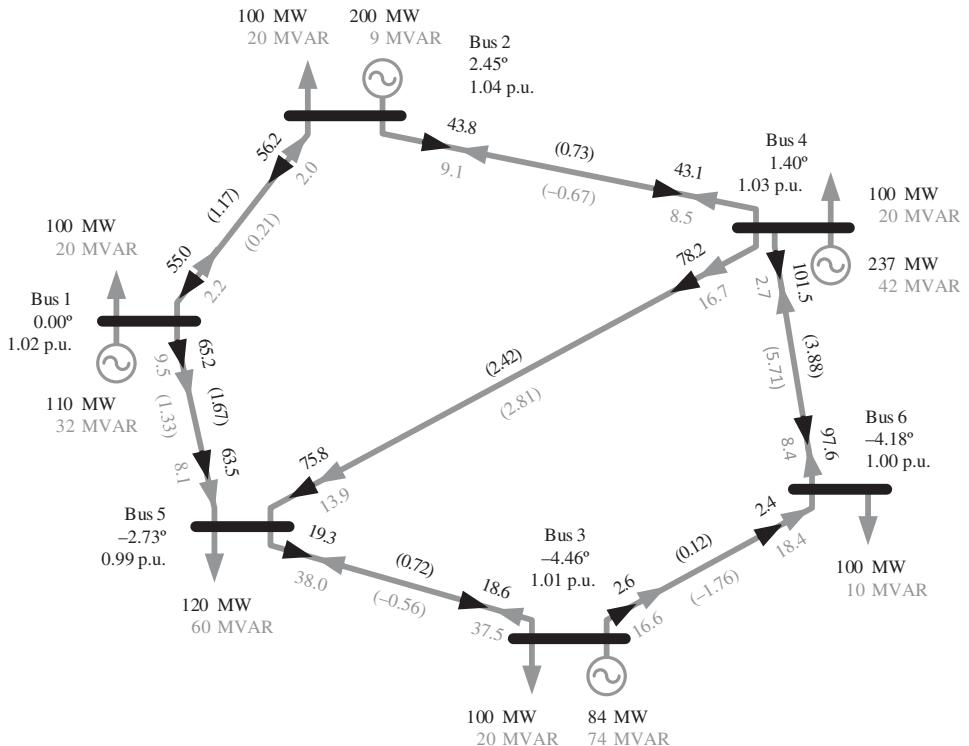


Figure 12.6 Modified six-bus power flow example.

The voltage magnitude at Bus 6 (a P, Q bus) is affected slightly (from 0.9951 to 0.9950 p.u.), although the change does not show up after rounding. In order to maintain the preset voltages at the P, V buses, reactive power generation increases at Buses 4 and 6, as it does at the slack bus. Indeed, system reactive generation now totals 157 MVAR, which are needed to accommodate the additional reactive load introduced at Bus 5 as well as a substantial increase in system reactive losses of 2.33 MVAR (up almost 50% from 4.74 to 7.07 MVAR). While real power is fixed at all buses other than 1 and 5, their voltage angles change as a result of the changed power flow pattern.

The most serious repercussion in this example occurs on Link 4–6, which was fully loaded before the change to Bus 5 was made. Owing to the vagaries of network flow, the change at Bus 5 results in a slight increase in real power flow from Bus 4 to Bus 6. Link 4–6 now carries slightly more current and apparent power, going from 100 to 101.5 MVA. This is significant because, in this hypothetical scenario, each transmission link has a thermal limit of 100 MVA. The power flow program thus shows the line becoming overloaded as a result of the change at Bus 5, even though Buses 4 and 6 are located on the opposite geographic end of the system, and neither generation nor load levels there were affected. In reality, this violation would mean that the proposed change is inadmissible and other options would have to be pursued—specifically, a generator other than Bus 1 would be required to increase generation—in order to meet the additional load without violating any transmission constraints.

The reader is encouraged to further tinker with power flow scenarios, for which the PowerWorld application is an invaluable tool. As this chapter illustrates, it is very difficult to produce an intuitive comprehension of power flow from the formal analytic description of the problem. Actual system operators, whose success hinges on just such an intuitive comprehension, develop it over the course

of time by empirical observation of countless scenarios (see Section 16.3.1). Experimentation with a small network model like the one just discussed offers students of power systems an opportunity both to create some contextual meaning for the abstract mathematical power flow relationships, and to appreciate the complexity and challenge of the task faced by system operators and engineers.

12.3.3 Conceptualizing Power Flow

Perhaps the most difficult conceptual aspect of a.c. power systems has to do with time. The synchronous oscillation along with its profile of voltage angles across the network provides a consistent temporal reference frame for an entire synchronous grid, which might span half a continent. The 60-cycle voltage oscillation is like a pulse that pervades this grid. Indeed, we can imagine a universal clock for the entire system, whose ticking marks the synchronized rhythm of all the connected generators. Yet, as seen in the power flow analysis, a time difference emerges and the oscillations do not coincide perfectly among the various buses. After arbitrarily assigning one bus as the reference for the system clock (a voltage angle of 0.00°), some buses will be slightly anticipating and others slightly dragging behind this pulse.¹¹ Converted into seconds, the time difference is minuscule: 1° is $1/360$ th of a cycle, which is $1/60$ th of a second, making 1° equal to about 46 microseconds (μs) or 0.0000463 seconds.

It could be tempting to attribute this time differential to the propagation time of a signal, but that would be incorrect. As mentioned in Chapter 1, transmission lines can easily be long enough so that the time it would take for an electrical signal to propagate at the speed of light from one end to the other is not negligible. Thus we might wonder about a delay between the voltage maximum (which could be considered a signal) occurring at different buses. Are the power-producing buses senders and the power-consuming buses receivers of a signal that is delayed by the distance between them?

Not exactly. The steady-state behavior we are concerned with does not correspond to the situation where a message originates from one distinct location to another at nearly the speed of light. A signal in the sense of “traveling message” in an a.c. grid would actually consist of a disturbance or departure from the 60-cycle oscillation in the background. Such disturbances (caused by faults, for example) do in fact propagate from one distinct location to another at the speed of light in a conductor, and when studying them, their “travel time” on transmission and distribution lines is meaningful (and can in fact be used to locate the disturbance source). These phenomena are distinct from power flow analysis, which deals with a steady operating state.

Initially establishing a steady state requires some pulse to make its rounds through the system as a disturbance signal—for example, while a section of the grid is energized by the first generator online, or as switches are closed between sections. But once these disturbances have echoed back and the transient effects have decayed, what is left is an ambient condition of being “energized” that resides everywhere in the system at once.

In fact, the time differentials expressed in the voltage-angle profile arise not from long-distance communication, but from the nature of a.c. power transfer as a strain across the transmission line. It is helpful to think of the angle difference like a *twist* in a rotating shaft that delivers mechanical power. Consider a *line shaft* as shown in Figure 12.7 that links numerous drive belts, pulleys or gears, commonly used in factories a century ago. A prime mover (such as a steam engine or water wheel) is driving the shaft forward, supplying mechanical power. Various machines are taking power from the shaft, by letting it drive their belts and thereby acting to hold back its rotation.

¹¹ By musical analogy, this is like a band playing syncopated notes that don't coincide on the downbeat, but each player remains consistent from one measure to the next.



Figure 12.7 Common rotating line shaft in a yarn spinning factory (Leipzig, ca. 1925) as a mechanical analog to a.c. power transfer. Source: Atelier Hermann Walter/Wikimedia Commons/Public Domain.

Clearly, maintaining a constant rotational speed of the shaft hinges on the balance of torque from the belt or gears driving it forward, and those being driven. This is analogous to maintaining a constant a.c. frequency, except that in the case of the electric grid, there are multiple drivers of the shaft interspersed with the loads.

Now, a rigid steel shaft will not exhibit much deformation in the process. But imagine a line shaft made of rubber. Then it becomes intuitive that in those places where the shaft is being driven forward, it will twist in the forward direction relative to the places where belts are holding back the rotation. This mechanical twist angle is an excellent analogy to the voltage phase angle.

If we had marked the rubber shaft in advance with a straight line as a reference (or perhaps stuck a row of pins into it along its length), we could then observe the twist at any location along the rotating shaft by noting that the line (or the local pin) crosses a reference position (say, vertical) at a slightly different time than its neighbors elsewhere along the shaft. Note that nothing material is traveling from one place to another, nor does it take time to do so. The twist is just a characterization of the deformation of the shaft in its steady state of rotation. As such, it is perfectly analogous to the local voltage phase angle.

The line shaft analogy also illustrates a key property of electric power that often challenges economists: namely, that there is no such thing as a package of goods being physically delivered from source to load. Rather, power is transferred into or out of the system locally by each generator or load as they interact with the energized medium through electromagnetic forces, where the power injected or withdrawn by a rotating machine can be expressed as torque times rpm. The transfer of power through the medium could be imagined in terms of a series of microscopic distortions along the shaft, but there is no unique packet of energy traveling. Rather, because the shaft is in a global state of rotation, it can be transacted with locally.

Likewise, the key property of the energized grid is that an alternating voltage is already present everywhere, like an ambient vibration. Therefore, it is more appropriate to think of power transfer as being perfectly instantaneous at every location. It would be incorrect physically to associate some specific power injected in one place with some specific power consumed elsewhere—despite the fact that electricity contracts are expressed as if this were the case. In this sense, our whole notion of

“power flow” is always on the verge of being wrong, if we attempt to identify particular quantities of power as though they were unique entities.

The elasticity of the rubber shaft is analogous to the impedance of a transmission line. This is consistent with the idea that a more elastic shaft will twist farther for a given amount of torque. Likewise, the voltage phase angle difference associated with some amount power transfer will be greater if there is a higher impedance, as seen in Eq. (7.1).

Further, we can think about dynamic behaviors. Imagine if a big engine driving our rubber shaft suddenly stops working. From the shock, the shaft will reverberate in a twist, with some oscillations as it slows down. This is analogous to the stability problem (see Section 13.4.4). Clearly, the bouncing behavior will get worse with a longer and more flexible shaft under high load—just like a long transmission line with high impedance.

The spinning rubber shaft is a great visualization aid for power transfer along a single transmission link. But now, what about an entire network? A different mechanical analogy imagines transmission lines as rubber bands tied together into a grid. Each bus is a place where the rubber bands are either suspended by hooks from the ceiling (generators) or have weights hanging from them (loads). The real power in megawatts injected or consumed at each node corresponds to the weight or amount of force pulling the node up or down. The voltage angle roughly corresponds to the elevation of each point in the rubber-band grid. The requirement that power injected equals power consumed corresponds to the requirement that the rubber-band grid be in balance, that is, neither fall down nor snap to the ceiling.

The rubber-band model visualizes dynamic stability in terms of what happens when a weight suddenly falls off (a load is lost) or a hook pops out of the ceiling (a generator goes offline). Even if we assume that the remaining hooks can accommodate the weight (i.e., generators compensate for the change in system load), the dynamic problem is that the network of rubber bands will bounce up and down following the sudden change. Thus, dynamic stability addresses the question of how much bouncing the hooks will tolerate before the whole web of rubber bands comes falling off.

Clearly, any given rubber band can only be stretched so far before it breaks. This translates into the observation that any given transmission link can only sustain a limited difference in voltage angle between its two ends (*steady-state stability*; see Section 13.4.3). Once this limit is exceeded and synchronicity is lost, the link no longer transmits power, just as the broken rubber band no longer transmits a force from weight to ceiling hook.

This analogy becomes a bit awkward if we try to stretch it further (so to speak) and bring line impedance into it. Rubber bands come in different elasticities and strengths, referring to the amount by which they stretch under a given tension, and their ability to withstand tension without breaking. A transmission line with a high impedance would correspond to a band that stretches farther under a given tension, either because it is longer or because it is more elastic. (We have a visualization problem here in that the dimension of linear distance in the rubber band model relates to voltage angle, not geographical distance.) To make the analogy work, we must require that all rubber bands “break” (i.e., violate a stability limit) when elongated by a certain number of inches. It would then hold true that the stability limit is increasingly important for longer lines and those with higher impedance. The thermal limit, by contrast, would be related to the amount of tension or force that can be sustained by each band, regardless of stretch.

A superconducting DC transmission line (Section 7.2.5) would translate into a perfectly strong and firm cord with no stretch at all. In the rotating shaft analogy, it would be perfectly rigid with no twist. If our electric grid were connected by such ideal links, nothing would bounce or stretch, and the subject of power flow analysis would become rather uninteresting.

This is about as far as the rotating shaft and rubber-band analogies go; trying to incorporate reactive power and voltage magnitude into these model is too contrived to be useful. Perhaps the most appropriate conclusion is that a.c. power systems have a certain complexity which, in its defiance of human intuition, is unmatched by any mechanical system.

12.4 Power Flow Equations and Solution Methods

12.4.1 Derivation of Power Flow Equations

In Section 12.2.2, we stated the known and unknown variables for each of the different types of buses in power flow analysis. The *power flow equations* show explicitly how these variables are related to each other.

The complete set of power flow equations for a network consists of one equation for each node or branch point in this network, referred to as a bus, stating that the complex power injected or consumed at this bus is the product of the voltage at this bus and the current flowing into or out of the bus. Because each bus can have several transmission links connecting it to other buses, we must consider the sum of power entering or leaving by all possible routes. To help with the accounting, we will use a summation index i to keep track of the bus for which we are writing down the power equation, and a second index k to keep track of all the buses connected to i .

We express power in complex notation, which takes into account the two-dimensionality—magnitude and time—of current and voltage in an a.c. system. As shown in Section 3.5.2, complex power S can be written¹² in shorthand notation as

$$S = VI^*$$

where all variables are complex quantities and the asterisk denotes the complex conjugate of the current.¹³

Recall that S represents the complex sum of real power P and reactive power Q , where P is the real and Q the imaginary component. At different times it may be convenient to either refer to P and Q separately or simply to S as the combination.

In the most concise notation, the power flow equations can be stated as

$$S_i = V_i I_i^*$$

where the index i indicates the node of the network for which we are writing the equation. Thus, the full set of equations for a network with n buses would look like

$$S_1 = V_1 I_1^*$$

$$S_2 = V_2 I_2^*$$

...

$$S_n = V_n I_n^*$$

¹² In this chapter, we reserve boldface notation for matrices.

¹³ The complex conjugate of a complex number has the same real part but the opposite (negative) imaginary part; see Section 3.3.5. It is used here to produce the correct relationship between voltage and current angle—their difference, not their sum—for purposes of computing power.

We can choose to define power as positive either going into or coming out of that node, as long as we are consistent. Thus, if the power at load buses is positive, that at generator buses is negative.

So far, these equations are not very helpful, since we have no idea what the I_i are. In order to mold the power flow equations into something we can actually work with, we must make use of the information we presumably have about the network itself. Specifically, we want to write down the impedances of all the transmission links between nodes. Then we can use Ohm's law to substitute known variables (voltages and impedances) for the unknowns (currents).

Written in the conventional form, Ohm's law is $V = IZ$ (where Z is the complex impedance). However, when solving for the current I , it is easier to use the admittance Y (where $Y = 1/Z$), so that Ohm's law becomes $I = VY$. This allows us to indicate the absence of a transmission link with a zero (for zero admittance), creating a *sparse* matrix for large systems that greatly facilitates computation.

The relationship $I = VY$ is what we wish to write down and substitute for every I_i that appears in the power flow equations. But now we face the next complication: the total current I_i coming out of any given node or bus is in fact the sum of many different currents going between bus i and all other buses that physically connect to i . We will indicate these connected buses with the index k , where k could include all buses in the network from 1 to n . In practice, fortunately, only a few of these will actually have links connecting to bus i . For any bus k that is not connected to i , $Y_{ik} = 0$.

For the current from node i to node k , we would generally write

$$I_{ik} = (V_k - V_i)Y_{ik}$$

where Y_{ik} is the admittance between i and k .

Suppose we are analyzing Bus 1, which is connected to Buses 2 and 3. (In this case $i = 1$, and $k = 1, 2$, and 3.) By Kirchhoff's current law, the net current injection at Bus 1 from generation and/or load must equal the net outflow of current into the network, thus:

$$I_1 = I_{12} + I_{13}$$

Writing this in terms of bus voltages and admittances, and rearranging terms, we get

$$\begin{aligned} I_1 &= (V_1 - V_2)Y_{12} + (V_1 - V_3)Y_{13} \\ &= V_1(Y_{12} + Y_{13}) - V_2Y_{12} - V_3Y_{13} \end{aligned}$$

Here it becomes convenient to change the sign of the branch admittance, which allows us to write the current as a sum of positive terms. By letting $y_{12} = -Y_{12}$, we get:

$$I_1 = V_1(Y_{12} + Y_{13}) + V_2y_{12} + V_3y_{13}$$

This will extend very nicely when we sum over all nodes k that could possibly be connected to node i , in tidy summation notation.¹⁴ This summation over the index k means accounting for all the current that is entering or leaving this one particular node, i , by way of the various links it has to nodes k . (For the complete system of power flow equations, we will consider every value of the index i so as to consider power flow for every bus.)

Thus,

$$S_i = V_i I_i^* = V_i \left(\sum_{k=1}^n y_{ik} V_k \right)^*$$

It is important to note that the $k = i$ term is included in the summation. The admittance y_{ii} is called the *self-admittance*. This self-admittance is defined as the (positive) sum of all the

¹⁴ The notation with the capital Greek sigma (for “sum”) indicates the sum of indexed terms, with the index running from the value below the sigma ($k = 1$) to the value above it ($k = n$).

admittances connected to the i th bus. In the above example, $y_{11} = Y_{12} + Y_{13}$. By including the term for $k = i$ and making the branch admittances negative, we avoid the need to write out voltage differences between buses.

Recall from Section 3.3.7, the complex admittance $Y = G + jB$, whose real part G is the *conductance* and whose imaginary part B is called *susceptance*. The admittances of all the links in the network can be summarized by way of an *admittance matrix* \mathbf{Y} , where the lowercase $y_{ik} = g_{ik} + jb_{ik}$ indicates the matrix element that associates nodes i with k . More about the admittance matrix in Section 12.4.2.

To complete our expression for complex power at the i th bus, we expand the y 's into g 's and b 's (noting that the complex conjugate gives a minus sign in front of the jb):

$$S_i = V_i \sum_{k=1}^n (g_{ik} - jb_{ik}) V_k^*$$

After rearranging terms to look more organized, we write the voltage phasors out in longhand, first as exponentials and then broken up into sines and cosines:

$$\begin{aligned} S_i &= \sum_{k=1}^n |V_i| |V_k| e^{j(\theta_i - \theta_k)} (g_{ik} - jb_{ik}) \\ &= \sum_{k=1}^n |V_i| |V_k| [\cos(\theta_i - \theta_k) + j \sin(\theta_i - \theta_k)] (g_{ik} - jb_{ik}) \end{aligned}$$

The term $(\theta_i - \theta_k)$ in this equation denote the *difference* in voltage phase angle between nodes i and k , where the minus sign came from having used the complex conjugate of I initially. It is often convenient to abbreviate it $(\theta_i - \theta_k) = \theta_{ik}$.¹⁵

The equation we now have for S_i entails the product of two complex quantities, written out in terms of their real and imaginary components. By cross-multiplying all the real and imaginary terms, we can separate the real and imaginary parts of the result S , which will be the familiar P and Q . Taking care with the sign of j^2 , we obtain:

$$\begin{aligned} P_i &= \sum_{k=1}^n |V_i| |V_k| [g_{ik} \cos(\theta_i - \theta_k) + b_{ik} \sin(\theta_i - \theta_k)] \\ Q_i &= \sum_{k=1}^n |V_i| |V_k| [g_{ik} \sin(\theta_i - \theta_k) - b_{ik} \cos(\theta_i - \theta_k)] \end{aligned} \tag{12.1}$$

The complete set of power flow equations for a network of n buses contains n such equations for S_i , or pairs of equations for P_i and Q_i . This complete set will account for every node and its interaction with every other node in the network.

12.4.2 The Bus Admittance Matrix

There are many possible ways of organizing information about the electrical connectivity and impedance characteristics of a network.

One basic choice is whether to describe the network in terms of impedances or admittances. As mentioned earlier, admittance has the advantage of having many smaller or zero values, since the vast majority of nodes in a network are not directly connected to each other. This creates a more sparse matrix that is easier and faster to manipulate, while the information it contains about the

¹⁵ This voltage phase angle difference is called δ_{12} elsewhere, but θ is more commonly used in the context power flow analysis.

network is the same. There are some calculations for which an impedance matrix is better suited than an admittance matrix; one important example is fault analysis. These techniques are beyond the scope of this text.¹⁶ We will limit ourselves to the admittance matrix here, as it is far more commonly used for power flow analysis.

Another choice is whether to describe network *branches* or *nodes*. In the branch description, voltages are understood as being across and currents through an individual branch, so that Ohm's law can be written for each branch individually. The impedances or admittances of individual branches are sometimes called *primitive*. They can be collected in a primitive impedance or admittance matrix.¹⁷ Such a compilation of branch impedances or admittances is distinct from the information about their connectivity, that is, which network branches actually meet at a node. Our choice here will be to summarize information by node.

The bus admittance matrix \mathbf{Y}_{bus} , also called *nodal admittance matrix*, contains information about both the connectivity of the network and the numerical values of the admittances connecting each pair of nodes. It is the inverse of the *bus impedance matrix* \mathbf{Z}_{bus} :

$$\mathbf{Y}_{\text{bus}} = \mathbf{Z}_{\text{bus}}^{-1}$$

Both of these matrices are symmetrical (i.e., values can be flipped about the main diagonal). The entries along the main diagonal characterize an individual node, and the off-diagonal elements characterize the connections between a respective pair of buses without any implied sense of directionality: $y_{ik} = y_{ki}$ and $z_{ik} = z_{ki}$. The diagonal elements of \mathbf{Y}_{bus} are called *self-admittances* and the off-diagonal elements are the *negative branch admittances*. These definitions, including the negative signs, account for the connectivity. They let us obtain currents throughout the network just by multiplying the nodal admittance matrix by all the nodal bus voltages, without having to take any explicit voltage differences:

$$\mathbf{I} = \mathbf{Y}_{\text{bus}} \mathbf{V}$$

The way to construct an admittance matrix for any arbitrary network is first to convert that network into a *Norton equivalent* circuit (see Section 2.5.2). This means expressing all elements as some combination of current sources and parallel admittances.¹⁸ If branches were characterized in terms of impedances (in ohms), these must be converted to admittances (in siemens). The generators and loads in the power system are represented by current sources in the Norton equivalent. These do not affect the admittances. When creating the admittance matrix for a circuit diagram, the current sources are ignored.

Note that in a conventional circuit diagram, we include a “zero” node corresponding to ground. This ground or reference node is not counted in the bus admittance matrix. In the standard one-line

¹⁶ An excellent reference for constructing and using the impedance matrix remains the classic text by J.J. Grainger and W.D. Stevenson, *Power System Analysis* (McGraw Hill, 1994).

¹⁷ The *primitive* or *branch admittance matrix* is commonly denoted by $[y]$. It captures only the admittance of each individual branch, and needs to be combined with an *adjacency* or *bus incidence matrix* A that contains information about the connectivity of the network along with reference directions for currents and voltages. For connoisseurs of linear algebra, $\mathbf{Y}_{\text{bus}} = A^T [y] A$. The reference directions are assigned in order to relate branch voltages to nodal current injections (i.e., defining current positive into or out of each branch at a given node). In graph theory, one would say that the network is represented by a *directed graph*. The reference direction is captured by assigning a +1 or -1 to each connection, depending on whether it is “entering” or “leaving” the node. Either convention works as long as it is applied consistently; in power systems, we conventionally assign +1 to a branch leaving a node. The adjacency matrix has zeroes along the main diagonal.

¹⁸ The corresponding procedure for an impedance matrix draws a Thévenin equivalent circuit, with a combination of voltage sources and series impedances.

diagram of a power system network, we don't explicitly show the ground; generators and loads are depicted as just dangling off their buses. The way to reconcile these representations is to imagine a ground plane behind the bus network diagram, to which all these dangling devices connect: this is simply stretching the circuit diagram into three dimensions, with the reference node out of view. Thus, when we speak of a “current injection” at Bus i , we are talking about current going from the ground plane into the network. Applying Kirchhoff's current law at every bus ensures that the overall net sum of currents to ground is zero. If we characterized the ground node or plane as its own circuit node, it would contain only redundant information. Therefore we simply omit it from the drawing of buses, and it does not get its own row or column in the admittance matrix.

To construct \mathbf{Y}_{bus} , we inventory the self-admittances for each node, by adding all admittances that are connected to that node. This includes the branch admittances to all adjacent nodes, as well as any shunt admittances associated with the individual node.¹⁹ The self-admittances become the diagonal matrix elements y_{ii} . The off-diagonal matrix elements $y_{ik} = y_{ki}$ will be the negative branch admittances.²⁰

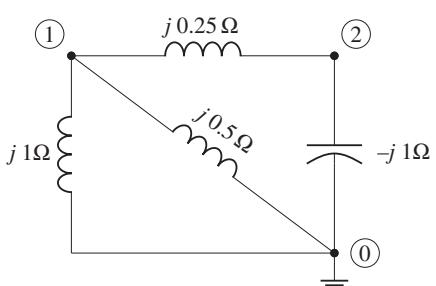
In the bus network diagram as used in power flow analysis, the current injections are implicit in loads and generation, which are specified in terms of power. These power injections also implicitly capture the admittances through loads and generators to ground, so we don't need to include them in the network description. This representation is an alternative to the more detailed description from the Norton equivalent circuit, where anything connected between a node and ground appears as a combination of current source and impedance.

Note that in all of these cases, it is possible to specify impedances and admittances in physical units of ohms and siemens, respectively, but especially for large networks it is most common to use per-unit quantities.

Example

Consider the network diagram in Figure 12.8. Nothing will happen on this circuit because it has no source—but we can still write an admittance matrix for it. In this example, we are given impedances, so these need to be converted. We first write the primitive admittances Y_{ik} for each

Figure 12.8 Simple network to illustrate obtaining branch admittances.



¹⁹ For example, in the medium transmission line model (Section 9.3.5), capacitances are divided in half, and each half is associated with one line end. These capacitances count toward the self-admittance, but not the branch admittance

²⁰ That negative sign comes out of the conversion from branch to nodal admittances, as detailed previously. It ensures that all voltage differences between nodes are accounted for correctly when we multiply the bus admittance matrix by nodal voltages to obtain nodal currents.

branch, and then the matrix elements y_{ik} ²¹:

$$Y_{10} = \frac{1}{j1} + \frac{1}{j0.5} = -j2$$

$$Y_{12} = \frac{1}{j0.25} = -j4$$

$$Y_{20} = \frac{1}{-j1} = j1$$

$$y_{11} = Y_{10} + Y_{12} = -j2 - j4 = -j6$$

$$y_{22} = Y_{12} + Y_{20} = -j4 + j1 = -j3$$

$$y_{12} = -Y_{12} = j4$$

The bus admittance matrix looks like this:

$$\mathbf{Y}_{\text{BUS}} = j \begin{bmatrix} -6 & 4 \\ 4 & -3 \end{bmatrix}$$

Note that the matrix has dimension 2×2 because one of the circuit nodes in the diagram is the ground reference.

Example

Consider the diagram in Figure 12.9, where two buses are connected by a transmission line modeled as a medium-length line, with series impedance $Z = j0.2$ p.u. and shunt admittance (due to capacitance) $Y = j0.2$ p.u. The capacitance is split in half and allocated to each end. The admittance matrix elements are as follows:

$$y_{11} = y_{22} = \frac{1}{Z} + \frac{Y}{2} = \frac{1}{j0.2} + j0.1 = -j5 + j0.1 = -j4.9$$

$$y_{12} = y_{21} = -\frac{1}{Z} = j5$$

The matrix looks like this:

$$\mathbf{Y}_{\text{BUS}} = j \begin{bmatrix} -4.9 & 5 \\ 5 & -4.9 \end{bmatrix}$$

When transmission line data are tabulated for input into power flow solvers, the shunt admittance is conventionally listed as B (indicating it is a susceptance only), in a separate column from the series components R and X . Very occasionally, there may be nonzero entries for shunt conductance G , which would similarly be split in half.

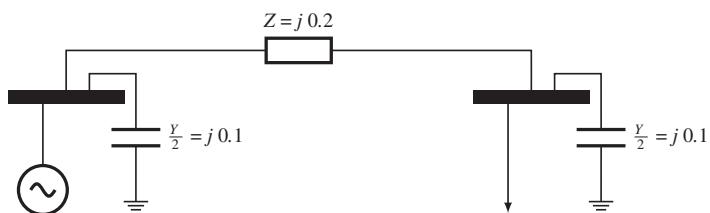


Figure 12.9 Simple two-bus network to illustrate allocation of transmission line capacitance.

²¹ Here we use lowercase letters for the elements of the bus admittance matrix and capital letters for the primitive values; some authors may use a reverse convention.

12.4.3 Solution Methods

There is no analytical, closed-form solution for the set of power flow equations given in Section 12.4.1. In order to solve the system of equations, we must proceed by a numerical approximation that is essentially a sophisticated form of trial and error.

To begin with, we assume certain values for the unknown variables. For clarity, let us suppose that these unknowns are the voltage angles and magnitudes at every bus except the slack, making them all P , Q buses (it turns out that having some P , V buses eases the computational volume in practice, but it does not help the theoretical presentation). In the absence of any better information, we would probably choose a *flat start*, where we assume the initial values of all voltage angles to be zero (the same as the slack bus) and the voltage magnitudes to be 100% of the nominal value, or 1 p.u. In other words, for lack of a better guess, we suppose that the voltage magnitude and angle profile across the system is completely flat.

We then plug these values into the power flow equations. Of course, we know they do not describe the actual state of the system, which was supposed to be consistent with the known variables (P 's and Q 's). Essentially, this will produce a contradiction: based on the starting values, the power flow equations will predict a different set of P 's and Q 's than we stipulated at the beginning. Our objective is to make this contradiction go away by repeatedly inserting a better set of voltage magnitudes and angles: as our voltage profile matches reality more and more closely, the discrepancy between the P s and Q s, known as the *mismatch*, will shrink. Depending on our patience and the degree of precision we require, we can continue this process to reach some arbitrarily close approximation. This type of process is known as an *iterative* solution method, where “to iterate” means to repeatedly perform the same manipulation.

The heart of the iterative method is to know how to modify each guess in the right direction and by the right amount with each round of computation (iteration), so as to arrive at the correct solution as quickly as possible. Specifically, we wish to glean information from our equations that tells us which value was too high, which was too low, and approximately by how much, so that we can prepare a well-informed next guess, rather than blindly groping around in the dark for a better set of numbers.

There are several standard techniques for doing this. The ones most commonly used in power flow analysis are the Newton–Raphson, the Gauss, and the Gauss–Seidel iterations. We introduce the basic idea of Newton's method in Section 12.4.4, and work through a step-by-step example with the Newton–Raphson algorithm. Once this process is understood conceptually, readers should be able to interpret alternative techniques presented in other references. In practice, the choice of algorithm for a particular situation involves a trade-off among the number of iterations required to arrive at the solution, the amount of computation required for each iteration, and the degree of certainty with which the solution is found.

Regardless of which method is used, we will need to press our power flow equations for the crucial information about the error in each iteration, to determine the next one. Some readers may recognize this as a kind of *sensitivity analysis*, which asks how much one variable is affected by changes in another. We obtain this information by writing down the *partial derivatives* of the power flow equations, or their rates of change with respect to individual variables. Specifically, we need to know the rate of change of real and reactive power, each with respect to voltage angle or magnitude. This yields four possible combinations of partial derivatives.²² For example, $\partial P / \partial \theta$ (read: “partial P partial theta”) is the partial derivative of real power with respect to voltage angle, and similarly there

²² The partial derivative means the rate of change of a function with respect to only one of several variables, and is conventionally indicated by the curly ∂ instead of plain d for “differential element.”

are $\partial P/\partial V$, $\partial Q/\partial \theta$, and $\partial Q/\partial V$.²³ Each of these combination is in fact a matrix (known as the Jacobian matrix) that, in turn, includes every bus combined with every other bus. In expanded form, with three buses, $\partial P/\partial \theta$ will look like this:

$$\frac{\partial P}{\partial \theta} = \begin{pmatrix} \frac{\partial P_1}{\partial \theta_1} & \frac{\partial P_1}{\partial \theta_2} & \frac{\partial P_1}{\partial \theta_3} \\ \frac{\partial P_2}{\partial \theta_1} & \frac{\partial P_2}{\partial \theta_2} & \frac{\partial P_2}{\partial \theta_3} \\ \frac{\partial P_3}{\partial \theta_1} & \frac{\partial P_3}{\partial \theta_2} & \frac{\partial P_3}{\partial \theta_3} \end{pmatrix}$$

Each of these four types of partial derivatives ($\partial P/\partial \theta$, $\partial P/\partial V$, $\partial Q/\partial \theta$, and $\partial Q/\partial V$) constitutes one partition of the big Jacobian matrix \mathbf{J} :

$$\mathbf{J} = \left[\begin{array}{c|c} \frac{\partial P}{\partial \theta} & \frac{\partial P}{\partial V} \\ \hline \frac{\partial Q}{\partial \theta} & \frac{\partial Q}{\partial V} \end{array} \right]$$

Within each partition there are rows for the θ or V and columns for the P or Q belonging to every bus (except the slack bus, so that the dimensionality of each partition is one less than the number of buses in the system).

We must now combine the system of equations and its partial derivatives with our guess for the unknown variables in such a way that it suggests to us a helpful modification of the unknowns, which will become our improved guess in the next iteration.

12.4.4 Iterative Computation

Our task can be stated as trying to find an unknown value of a function whose explicit form we do not know, based on information from elsewhere along the function. This can be done with a *Taylor series expansion*, which may be familiar to readers who have studied calculus. The idea is that we can express the unknown value of the function $f(x)$ at some particular x in terms of two pieces of information: the function's value at a different, nearby x ; and the rate of change of the function with respect to x —its slope—at the same nearby x . Suppose we already know the value f at location x , and we also know how steep the function is there. Now we want to learn the value of f at the nearby location that we call $x + \Delta x$, so that Δx represents the difference between the two x 's. If the function $f(x)$ is a straight line, we can write

$$f(x + \Delta x) = f(x) + f'(x)\Delta x$$

where $f'(x)$ is the *first derivative* or slope of the line at location x .

In the more general case, where $f(x)$ is not a straight line, but some type of curve, this equation is incomplete; we would have to include additional higher-order terms that correct for the curvature. Specifically, we would include the second derivative (the rate of change of the rate of change, which describes the upward or downward curvature) to correct the straight-line approximation, multiplied once again by the increment Δx . We may also need to include the third derivative or more, depending on how curvy the function is. Each n th term gets successively scaled down by a factor of $n!$ (n factorial). Note that this procedure applies only to functions that are infinitely differentiable; in other words, they cannot have corners, spikes or discontinuities.

²³ Note that when we write V we mean the magnitude of V , which would be more properly designated by $|V|$ except that the notation is already cumbersome enough without the absolute value signs.

Written out, the Taylor series looks like

$$f(x + \Delta x) = f(x) + f'(x)\Delta x + \frac{f''(x)}{2!}\Delta x^2 + \frac{f'''(x)}{3!}\Delta x^3 + \dots$$

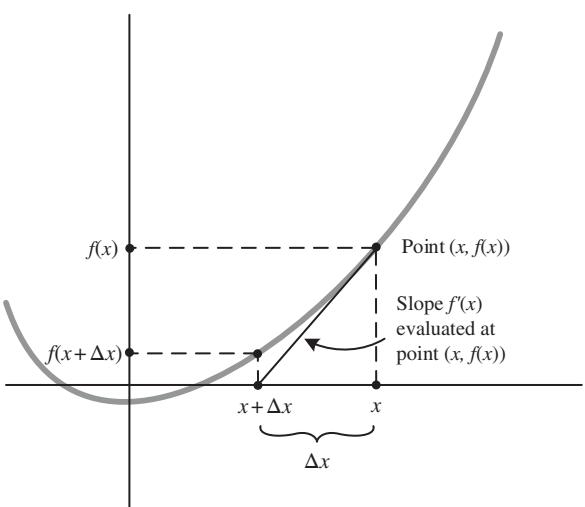
up to the point where the function's higher-order derivatives are zero.

Fortunately, we are in the business of making approximations rather than finding exact values. If the Δx is not too big and the function not too radically curved, then the higher-order terms ought to be quite small compared to the first derivative term. Thus, we can use the linear version for a general function, with the understanding that it will not take us exactly to the new value $f(x + \Delta x)$, but a good bit of the way there and almost certainly closer than we were before.²⁴

The reader may wonder why we don't simply plug the new x -value, $x + \Delta x$, into the function $f(x)$. Usually, this is because we do not know how to write down $f(x)$ or $f'(x)$ in algebraic form, even though we have numerical values for f and f' at some particular x . In our application, we happen to know what $f(x)$ looks like algebraically—the power flow equations—but we cannot solve it in the way that we would like. Specifically, we would like to go backwards to find the x that will yield a particular value of f . Even though we know how to get f from any given x , we can't simply solve for x given the f because the equation does not allow itself to be turned inside out. Specifically, given the θ 's and V 's in the power flow equations, we can readily solve for the P 's and Q 's, yet we cannot go backwards from the P 's and Q 's to explicitly solve for the θ 's and V 's. Thus, we are forced to try out different sets of θ 's and V 's (represented by the x 's) until we hit the right P 's and Q 's (represented by the f).

The standard way to proceed is to rearrange the equations as necessary so that the target value is $f(x) = 0$. The problem can then be stated in the tidy format, “Find the x that makes $f(x) = 0$ a true statement.” It is illustrated in Figure 12.10, where we start with a certain x and a known value $f(x)$ that is not zero, but wish to find the Δx for which the value of the function $f(x + \Delta x)$ is zero. Note that in this diagram Δx happens to be negative (i.e., measuring to the left), but it could go either way. After writing down the first terms of the Taylor series and declaring that $f(x + \Delta x) = 0$, it takes

Figure 12.10 Newton's method.



24 This is true unless the function is very badly behaved or we started in an awkward spot.

only a minor manipulation to solve explicitly for the Δx that makes this true:

$$f(x + \Delta x) = f(x) + f'(x)\Delta x$$

$$0 = f(x) + f'(x)\Delta x$$

$$\Delta x = -f(x)/f'(x)$$

We have then found the Δx that must be added to the original x to obtain the new x -value, for which the function is zero.

But because the function is curved, not straight, our answer will not be exactly right. We have evaluated the derivative $f'(x)$ at the location of our original x , meaning that we have used the slope at that location to extrapolate where the function is going. But in reality, the function's slope may change along the way. The higher-order terms of the Taylor series would address this problem, but they contain awkward squares and cubes. Instead of dealing with such terms, we simply repeat the linear process: we use the new location, $x + \Delta x$, as our starting point for another iteration. Since $x + \Delta x$ is presumably much closer to our target than the original x —which we can verify by checking that $f(x + \Delta x)$ is closer to zero than $f(x)$ —the next time it should be easier to get even closer, with a smaller Δx . Depending on the precision we desire in getting f to zero, we can repeat the process again with more iterations after that, or call it a day. This, in essence, is Newton's method for finding the zero-crossing of a function.

In any case, we will probably have many x 's and Δx 's around and ought to keep track of which iteration they belong to. One way to label them is with a superscript like x^v , where v (Greek lower-case nu) stands for the number of iterations (x^1, x^2, \dots and $\Delta x^1, \Delta x^2, \dots$) and is not to be mistaken for an exponent. The process of approaching a value of x for which $f(x) \approx 0$ is illustrated in Figure 12.11 for two iteration steps. Clearly, the more the slope changes between x^0 and the solution (i.e., the more the function is curved), the more steps will be required to get close. Based on the diagram, the

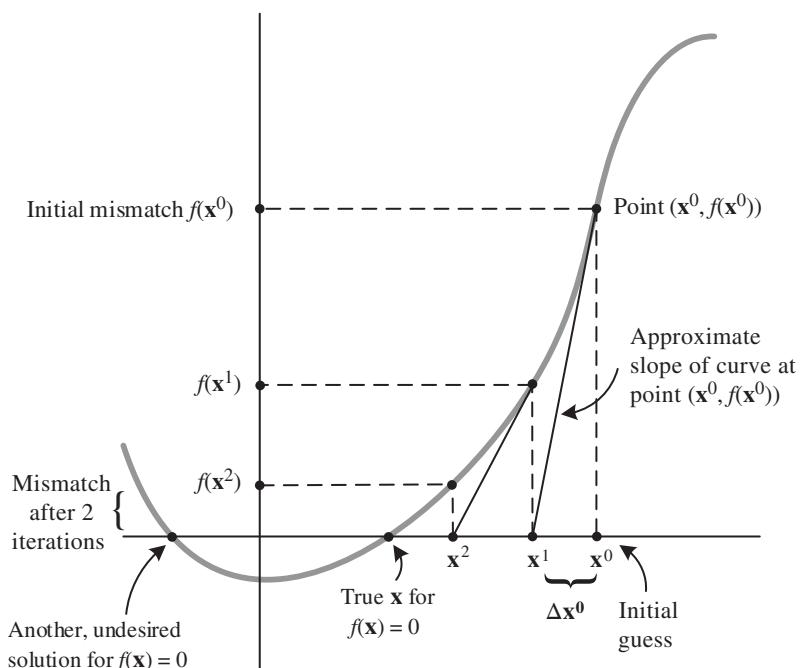


Figure 12.11 Iterative process of approximating $f(x) = 0$.

reader can also visualize how this approximation process can still succeed even if the slope of the line drawn to choose the next x^v is not precisely equal (but bears a reasonable resemblance) to the actual slope of the curve; this property is used in shortcuts such as the dishonest Newton-Raphson method discussed later in this chapter that avoid some of the tedious computation of the exact derivatives.

The Jacobian matrix is essentially a large version of the derivative $f'(x)$ used in Newton's method, with multiple f 's and x 's tidily summarized into the single, bulky object labeled \mathbf{J} . The function $f(x)$ itself captures the power flow equations $P(\theta, V)$ and $Q(\theta, V)$. We write $\mathbf{f}(x)$ in boldface to indicate that it is a vector containing an organized set of several numbers (one P and one Q for each bus except the slack). However, in order to keep with the format of searching for $f(x) = 0$, we define $\mathbf{f}(x)$ as the *difference* between the $P(\theta, V)$, $Q(\theta, V)$ computed from the power flow equations as a function of x , and the P and Q injections at each of the buses that are given at the outset. In other words, $\mathbf{f}(x)$ represent the *mismatch*, which we want to get as close to zero as reasonably possible.

In principle, we are now able to combine all the information at our disposal into improving our guess for the θ 's and V 's (the x 's). Let us label our initial guess for θ and V as \mathbf{x}^0 , where the boldface \mathbf{x} indicates the vector composed of one θ and one V for each bus (except the slack) and the superscript 0 indicates the zeroth iteration. Now we might simply adapt the expression for Δx from Newton's method (or some variation that corresponds to other solution methods, though the basic idea is always the same),

$$\Delta x = -f(x)/f'(x)$$

and substitute our matrix and vector quantities

$$\Delta \mathbf{x} = -\mathbf{f}(\mathbf{x})/\mathbf{J}$$

But stop! We've just made every mathematician cringe, because dividing by a matrix is not something one does. We need to use the proper *inverse* of the matrix, \mathbf{J}^{-1} , and write

$$\Delta \mathbf{x} = -\mathbf{J}^{-1}\mathbf{f}(\mathbf{x})$$

The inverse of a matrix is obtained by a tedious but tractable procedure in linear algebra, which quickly grows more cumbersome with increasing size of the matrix.²⁵ In the days of paper and pencil, half a dozen rows and columns would have easily defeated the most diligent scribe, and iterative power flow solution for large a.c. networks was simply not an option. Modern computing allows us to solve systems with hundreds and even thousands of buses, where inverting the Jacobian matrix is the critical part of the computational effort. Reasonably sized matrices are easily inverted today with calculators, Matlab or online tools—a capability still best appreciated through the character-building experience of solving a small system by hand.

Having obtained a correction $\Delta \mathbf{x}$ by hook or by crook, we add it to the old \mathbf{x} in order to proceed to the next iteration. Specifically, we write:

$$\mathbf{x}^{(v+1)} = \mathbf{x}^{(v)} + \Delta \mathbf{x} = \mathbf{x}^{(v)} - \mathbf{J}(\mathbf{x}^{(v)})^{-1}\mathbf{f}(\mathbf{x}^{(v)})$$

where the superscript v indicates the iteration number, with optional parentheses to prevent mistaking it for an exponent.²⁶

²⁵ Analogous to the inverse of a scalar, which gives 1 when multiplied by the original number, the inverse of a matrix produces the *identity matrix* when multiplied with the original matrix. Note that rearranging the equation to read $\mathbf{J}\Delta \mathbf{x} = -\mathbf{f}(\mathbf{x})$ does not remove the awkwardness because we must still solve for $\Delta \mathbf{x}$.

²⁶ We will sometimes drop the parentheses for convenience. Other notation options are subscripts or simply $x(v)$ to denote iteration count.

With successive iterations, the value of $\mathbf{f}(\mathbf{x}^{(v)})$ should get smaller. When it reaches zero—or close enough, according to a chosen convergence threshold—it means that there is no more mismatch. The $\mathbf{x}^{(v)}$ at that point (i.e., the θ 's and V 's from the v th iteration) give us the operating state of the power system that is consistent with the P 's and Q 's we specified initially. We may need to verify, though, that $\mathbf{x}^{(v)}$ is a realistic and true solution for our power system, as opposed to some mathematical fluke, which can occur when a function has more than one zero-crossing.

We have now found θ and V for each P, Q bus. For any P, V buses, we would have found θ and Q as part of our computational process. There are several steps left to produce the complete output of the power flow analysis. First, by writing the power flow equation for the slack bus, we determine the amount of real power generated there. This tells us how many MW of losses there are in the system, as we can now compare the total MW generated to the total MW of load demand. Also, by using Ohm's law for every transmission link, we solve explicitly for each line flow, in amperes or MVA. Through the θ 's and V 's at each bus, we have information about the real and reactive power both going into and coming out of each link, and by subtracting we can specify the real and reactive losses on each link. Finally, we format the output and compare it to external constraints such as line flow limits so as to flag any violations. We have now completely described the system's operating state based on a given generation dispatch and combination of loads.

12.4.5 Power Flow Example

For illustration, let us work through the smallest possible example, a two-bus power system, with the Newton–Raphson method.²⁷ In this tiny system, there is only one bus to be solved for by iteration, and all the arithmetic is tractable by hand.

Figure 12.12 illustrates the case. We have Bus 1 as the slack bus, whose voltage is set at $V_1 = 1.00\angle 0^\circ$ p.u. We don't know yet how much power the generator at Bus 1 needs to inject. What we do know is that the load at Bus 2 is demanding exactly 2.0 p.u. of real and 1.0 p.u. of reactive power (after choosing $S_{\text{BASE}} = 100$ MVA); that there is an impedance of $Z = j0.1$ p.u. between Bus 1 and 2; and that there will be some reactive but no resistive losses (since the line impedance is purely imaginary). We don't know what the voltage at Bus 2 will have to be in order to deliver the demanded power, or how much current has to flow.

Constructing a bus admittance matrix \mathbf{Y}_{bus} seems overkill for this example, since it is built on just a single value, the impedance $Z = j0.1$ between Bus 1 and 2. But creating the proper matrix with self- and mutual admittance terms is a useful exercise and reminder of where the minus signs go. Per the procedure introduced in Section 12.4.2, it is given by:

$$\mathbf{Y}_{\text{bus}} = \begin{bmatrix} y_{11} & y_{12} \\ y_{21} & y_{22} \end{bmatrix} = \begin{bmatrix} -j10 & j10 \\ j10 & -j10 \end{bmatrix}$$

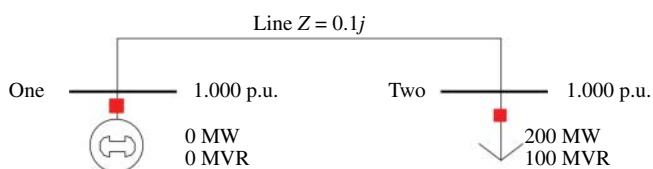


Figure 12.12 Two-bus power flow example visualized in *PowerWorld*. Source: U.S.-Canada Power System Outage Task Force, 2004/United States Department of Energy/Public Domain.

²⁷ This teaching example was created by Tom Overbye, author of *PowerWorld*TM. A free educational version of *PowerWorld*TM is downloadable from www.powerworld.com.

Using the general power flow equations (12.1) with $i = 2$, the net real and reactive power injections at Bus 2 are given by

$$P_2 = \sum_{k=1}^2 |V_2| |V_k| [g_{2k} \cos(\theta_2 - \theta_k) + b_{2k} \sin(\theta_2 - \theta_k)]$$

$$Q_2 = \sum_{k=1}^2 |V_2| |V_k| [g_{2k} \sin(\theta_2 - \theta_k) - b_{2k} \cos(\theta_2 - \theta_k)]$$

where the summation index k will take on the values 1 and 2.

We will now write equations for net power at each bus to be solved for (only Bus 2 in this example), which are the entries of the mismatch function $\mathbf{f}(\mathbf{x})$. When we have found the correct \mathbf{x} , this mismatch will approach zero, stating that the power injected at each bus by generation or load (as given in the problem statement) equals the power injected from that bus into the network by way of the power flow equations (as calculated from \mathbf{x}). Note that a positive power injection corresponds to generation, and power flow *away from* that bus on the transmission lines. Loads represent a negative injection. Since we subtract the injections from the power flows, loads appear as positive terms in the mismatch.

Thus, using the given values of $|V_1| = 1.0$, $g_{ik} = 0$ (since the line is purely inductive), $b_{21} = y_{21} = j10$, $P_2 = -2.0$ and $Q_2 = -1.0$, and setting net power equal to zero, we get:

$$\begin{aligned} P_{2\text{net}} &= |V_2|(10 \sin \theta_2) + 2.0 = 0 \\ Q_{2\text{net}} &= |V_2|(-10 \cos \theta_2) + |V_2|^2(10) + 1.0 = 0 \end{aligned} \quad (12.2)$$

where all the terms with g_{ik} have dropped out. Also, in the expression for P_2 , the $k = 2$ term disappears because $\sin(\theta_2 - \theta_2) = 0$. To set up a power flow solution by Newton–Raphson for the two bus example, we define a solution vector \mathbf{x} for the voltage at Bus 2:

$$\mathbf{x} = \begin{bmatrix} \theta_2 \\ |V_2| \end{bmatrix}$$

It is standard practice to always express θ in units of radians within the power flow calculation.

Before we choose a starting value and begin to perform any iterations, let us write out the Jacobian matrix whose elements are the partial derivatives of the power flow equations (12.2) with respect to the two components of \mathbf{x} :

$$\mathbf{J}(\mathbf{x}) = \begin{bmatrix} \frac{\partial P_2}{\partial \theta_2} & \frac{\partial P_2}{\partial |V|_2} \\ \frac{\partial Q_2}{\partial \theta_2} & \frac{\partial Q_2}{\partial |V|_2} \end{bmatrix} = \begin{bmatrix} 10|V_2| \cos \theta_2 & 10 \sin \theta_2 \\ 10|V_2| \sin \theta_2 & -10 \cos \theta_2 + 20|V_2| \end{bmatrix}$$

We will evaluate the four elements of this Jacobian for each successive iteration of \mathbf{x} . For our zeroth iteration, we choose the standard *flat start*, zero angle and 1 p.u. magnitude:

$$\mathbf{x}^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

As our first calculation step, we plug these values for θ and $|V|$ into the power flow equations and into the Jacobian matrix:

$$\mathbf{f}(\mathbf{x}^{(0)}) = \begin{bmatrix} P_{2\text{net}} \\ Q_{2\text{net}} \end{bmatrix} = \begin{bmatrix} |V_2|(10 \sin \theta_2) + 2.0 \\ |V_2|(-10 \cos \theta_2) + |V_2|^2(10) + 1.0 \end{bmatrix} = \begin{bmatrix} 2.0 \\ 1.0 \end{bmatrix}$$

These numbers look just like the load P_L and Q_L , because the flat start implies that no power is transferred at all (since the voltage magnitude and angle are identical between Bus 1 and 2). Since

we want $P_{2\text{net}}$ and $Q_{2\text{net}}$ to converge to zero, we have a long way to go. Evaluating the Jacobian for $\theta_2 = 0$ and $|V_2| = 1.0$, we get:

$$\mathbf{J}(\mathbf{x}^{(0)}) = \begin{bmatrix} 10|V_2|(\cos \theta_2) & 10 \sin \theta_2 \\ 10|V_2|(\sin \theta_2) & -10 \cos \theta_2 + 20|V_2| \end{bmatrix} = \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}$$

Now we are ready to put it all together and compute the next iteration $\mathbf{x}^{(1)}$. Recall that we must invert the Jacobian to compute the correction term for pointing us in the best direction toward the new \mathbf{x} that will shrink $\mathbf{f}(\mathbf{x})$.

$$\mathbf{x}^{(v+1)} = \mathbf{x}^{(v)} - \mathbf{J}(\mathbf{x}^{(v)})^{-1} \mathbf{f}(\mathbf{x}^{(v)})$$

Plugging in the values from the $v = 0$ iteration, we get:

$$\mathbf{x}^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 10 & 0 \\ 0 & 10 \end{bmatrix}^{-1} \begin{bmatrix} 2.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} - \begin{bmatrix} 0.1 & 0 \\ 0 & 0.1 \end{bmatrix} \begin{bmatrix} 2.0 \\ 1.0 \end{bmatrix} = \begin{bmatrix} -0.2 \\ 0.9 \end{bmatrix}$$

A quick sanity check suggests that a small negative voltage angle θ_2 and a voltage magnitude $|V_2|$ a bit less than 1.0 both make sense, because real and reactive power are flowing from Bus 1 to Bus 2. With this encouraging news, we repeat the procedure of evaluating $\mathbf{f}(\mathbf{x})$ and $\mathbf{J}(\mathbf{x})$, now at $\mathbf{x}^{(1)}$:

$$\mathbf{f}(\mathbf{x}^{(1)}) = \begin{bmatrix} 0.9(10 \sin(-0.2)) + 2.0 \\ 0.9(-10 \cos(-0.2)) + 0.9^2(10) + 1.0 \end{bmatrix} = \begin{bmatrix} 0.212 \\ 0.279 \end{bmatrix}$$

We're pleased to note that $\mathbf{f}(\mathbf{x})$ is shrinking. For the Jacobian evaluated at the new \mathbf{x} , we have:

$$\mathbf{J}(\mathbf{x}^{(1)}) = \begin{bmatrix} 9(\cos(-0.2)) & 10 \sin(-0.2) \\ 9(\sin(-0.2)) & -10 \cos(-0.2) + 18 \end{bmatrix} = \begin{bmatrix} 8.82 & -1.986 \\ -1.788 & 8.199 \end{bmatrix}$$

That produces the following for the next iteration:

$$\mathbf{x}^{(2)} = \begin{bmatrix} -0.2 \\ 0.9 \end{bmatrix} - \begin{bmatrix} 8.82 & -1.986 \\ -1.788 & 8.199 \end{bmatrix}^{-1} \begin{bmatrix} 0.212 \\ 0.279 \end{bmatrix} = \begin{bmatrix} -0.233 \\ 0.8586 \end{bmatrix}$$

This time, the corrections to \mathbf{x} were more modest, and evaluating $\mathbf{f}(\mathbf{x})$ already gets us fairly close to zero:

$$\mathbf{f}(\mathbf{x}^{(2)}) = \begin{bmatrix} 0.0145 \\ 0.0190 \end{bmatrix}$$

To take it one more step, the reader is invited to evaluate the Jacobian at $\mathbf{x}^{(2)}$ and confirm that the next iteration comes to

$$\mathbf{x}^{(3)} = \begin{bmatrix} -0.236 \\ 0.8554 \end{bmatrix}$$

which yields

$$\mathbf{f}(\mathbf{x}^{(3)}) = \begin{bmatrix} 0.0000906 \\ 0.0001175 \end{bmatrix}$$

This tiny mismatch should fall below a reasonable error threshold, which means that we can consider $\mathbf{x}^{(3)}$ our solution. Converting θ_2 from radians into degrees, we have solved for

$$V_2 = 0.8554 \angle -13.52^\circ \text{ p.u.} = 0.832 + j0.200 \text{ p.u.}$$

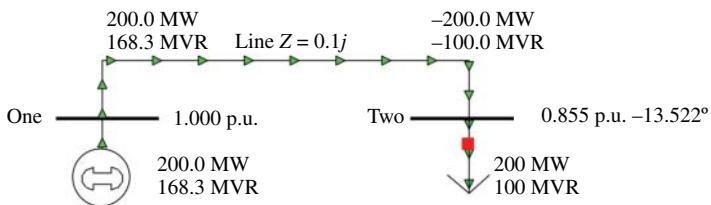


Figure 12.13 Two-bus power flow example solved in *PowerWorld*™.

This solution for the complex bus voltage fully and unambiguously specifies the operating state of the system, since we already know V_1 . With this information, it is straightforward to calculate all the other system values.

From Ohm's law, the current on the transmission line is

$$I = \frac{(V_1 - V_2)}{Z} = \frac{0.168 - j0.2}{j0.1} = 2 - j1.68 = 2.61\angle -40.0^\circ$$

and therefore the power injection at the slack bus is

$$S = I^*V = 2.61\angle 40.0^\circ \cdot 1.0\angle 0^\circ = 2.61\angle 40.0 = 2.00 + j1.68 \text{ p.u.}$$

With $S_{\text{BASE}} = 100 \text{ MVA}$, this gives

$$P_1 = 200 \text{ MW} \quad \text{and} \quad Q_1 = 168 \text{ MVAR}$$

for the slack bus real and reactive power injections, as illustrated in Figure 12.13. As expected, there are no real losses on this purely inductive transmission line, but the reactive losses are significant. The generator at the slack bus must provide the additional 68 MVAR of reactive power, or else the voltage at Bus 2 could not be maintained.

12.4.5.1 Low-voltage Solution

Although the numerical example above is about as simple as could possibly be contrived, it illustrates the diabolical nonlinearity of the power flow problem: specifying the complex power at Bus 2 does not unambiguously specify the state of the system! In fact, there are two mathematical solutions for V_2 that are equally consistent with the same power values at Bus 2. In the Newton–Raphson method, our choice of starting value $\mathbf{x}^{(0)}$ determines which solution the algorithm will converge to.

The flat start with $|V| = 1.0 \text{ p.u.}$ and $\theta = 0^\circ$ is generally a safe bet, because any solution in that neighborhood is likely to be physically reasonable and operationally stable. It is a conservative assumption that amounts to saying, let's suppose the power flows in the system are quite small, and the voltage differences between buses are not too dramatic.

In the above example, a starting guess with a small voltage magnitude (less than about 0.5) or a large angle (in the tens of degrees) will make the Newton–Raphson algorithm converge to the low-voltage solution $V_2 = 0.26\angle -49.9^\circ \text{ p.u.}$ As illustrated in Figure 12.14, this solution involves ridiculously high line losses—in fact, the transmission line becomes the major (inductive) load and sustains the majority of the voltage drop. In mathematical terms, the stipulated power condition at Bus 2 can be equally well met with low voltage and high current, but it would be unrealistic to actually operate the system in this condition in practice (see Section 13.4.5).

To further illustrate the complexity hiding within this toy example, we can map the fate of various starting guesses for \mathbf{x} . Figure 12.15 shows the *convergence regions*, where starting values in the dark gray (red in original) region will converge to the high-voltage solution, and the light gray (yellow)

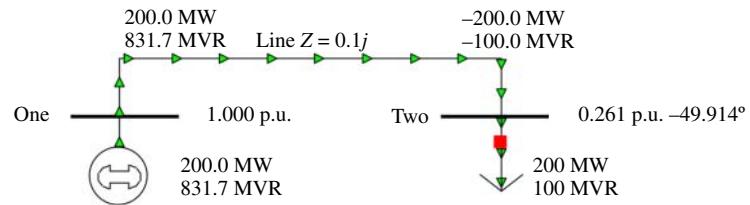


Figure 12.14 The unrealistic low-voltage solution to the two-bus power flow example.

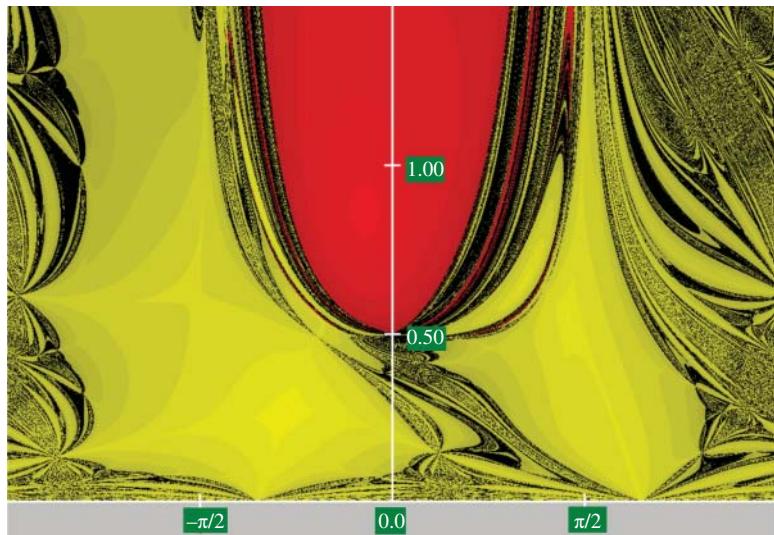


Figure 12.15 Convergence regions for the two-bus power flow example. Source: Courtesy of Tom Overbye.

region to the low-voltage solution. The shading within each region indicates how many iteration steps it takes to converge, with the lightest areas centering on the actual solutions. The dark bands are starting values for which the algorithm will be particularly challenged and tend to bounce back and forth through many iterations. The fractal boundary between the red and yellow regions is especially fascinating: even this simple example exhibits the *chaotic* property, where an arbitrarily small difference in the initial condition can have a dramatic effect on the outcome.²⁸ And especially in larger, more realistic systems, the solution algorithm may just fail to converge altogether.

Different iterative solution methods have different strengths and weaknesses. The Newton-Raphson method is most popular because it tends to converge relatively fast, and the number of iterations is independent of the number of buses in the network. It also does not matter which bus is selected as the slack bus. However, as we just saw, the results can be very sensitive to the initial guess.

The Gauss-Seidel method, by contrast, iterates on variables sequentially rather than all at once. It is based on Gaussian elimination and substitution, a routine but tedious procedure in linear algebra. There is no Jacobian matrix for the entire system to invert, so each iteration is computationally fast, but many iterations may be needed, and that number increases with the number of buses and branches in the network. The Gauss-Seidel method is less sensitive to the

²⁸ For further reading and inspiration, the classic and gorgeously illustrated hardcover *The Beauty of Fractals* by Heinz-Otto Peitgen and Peter Richter (Springer, 1986) is highly recommended.

initial guess for bus voltages, but more sensitive to the choice of slack bus. Gauss–Seidel sometimes converges where Newton–Raphson does not, and *vice versa*.

12.4.6 Shortcuts

12.4.6.1 Dishonest Newton–Raphson

As noted earlier, inverting the Jacobian matrix, which is used in the iterative procedure to find a better $x + \Delta x$, is the most computation intensive aspect of the Newton–Raphson method. One way to considerably reduce the computation volume and thereby speed up the process is simply not to update the Jacobian with every iteration step. This is known as the “dishonest” or “lazy” Newton–Raphson method. Of course, without using the correct derivatives, we may not be pointed toward the best Δx in each iteration. This brings some risk of making a very bad choice for the next x , but usually it just means it will take more iteration steps to converge.

We can visualize the dishonest Newton–Raphson method in Figure 12.11 as using the slope of $f(x)$ evaluated at x^0 to find not only the next adjustment Δx^0 , but reusing the same slope to choose Δx^1 and even subsequent iterations. For the curve in this figure, this would produce an x^2 not quite as close to the true x , but still progressing in the right direction. The computational savings from not repeatedly inverting a matrix are often worth the additional iteration steps.

Note that the accuracy of the solution is not affected by this shortcut, as we would choose the same error threshold for the mismatch; the method will either converge or it won’t. Using an outdated version of the Jacobian carries some risk, though, of derailing the solution progress and causing it to fail to converge, or sending it so far astray that it converges to the wrong solution.

Although this section is titled “shortcuts,” “clever detours” might be a better analogy. Using an inaccurate Jacobian is a bit like giving driving directions to the same destination along a detour route with fewer stop signs: we expect that even though it is less direct, it will be faster. Also, we hope that the driver won’t get lost.

12.4.6.2 Decoupled Power Flow

Instead of neglecting to update the Jacobian altogether, another way to save computation time is to use a simpler, approximate version of it that is more sparse (i.e., it contains many zero entries) and therefore faster to invert. Specifically, we may choose to decouple real power from voltage magnitude and reactive power for voltage angle, based on the common rule that real power flow relates mainly to differences in voltage angle, and reactive power mainly to differences in voltage magnitude. This rule derives from the key assumption that the reactive properties of transmission lines tend to outweigh their resistance, and also from the assumption that angle differences tend to be small.

Mathematically, we are claiming that two of the four distinct partitions of the Jacobian matrix, which contain the partial derivatives of P or Q with respect to θ and V , can be safely neglected. Without inserting any numerical values, we can examine the mathematical form of these partial derivatives and conclude which terms ought to be large and which ought to be small, based on the two assumptions just cited. What we will find is that the dependence of real power on voltage angle, $\partial P / \partial \theta$, ought to be substantial, while the dependence of real power on voltage magnitude, $\partial P / \partial V$, ought to be much smaller by comparison. For reactive power, we will find the opposite: $\partial Q / \partial \theta$ ought to be small, but $\partial Q / \partial V$ should be substantial.

Let’s walk through the process of examining the derivatives for a sample bus pair, 2 and 3 (to be general, we would write i and k). Since we are interested in power flow from one bus to another, we will consider only the derivatives with unequal indices (such as $\partial P_2 / \partial \theta_3$, as opposed to $\partial P_2 / \partial \theta_2$).

First, we write out the derivatives $\partial P_2 / \partial V_3$ and $\partial Q_2 / \partial \theta_3$, which we will show to be small:

$$\begin{aligned}\frac{\partial P_2}{\partial V_3} &= |V_2| [g_{23} \cos(\theta_2 - \theta_3) + b_{23} \sin(\theta_2 - \theta_3)] \\ \frac{\partial Q_2}{\partial \theta_3} &= |V_2| |V_3| [g_{23} \cos(\theta_2 - \theta_3) + b_{23} \sin(\theta_2 - \theta_3)]\end{aligned}$$

We now observe the implications of our two assumptions. If a transmission link's reactive effects substantially outweigh its resistive effects, this means its conductance g_{23} is a much smaller number than its susceptance b_{23} .²⁹ This makes the cosine terms small, as they are multiplied by the g's. The sine terms are multiplied by the b's, so they could be substantial based on that consideration. However, the sine terms are also small, for a different reason: if the voltage angle difference $\theta_2 - \theta_3$ between buses is small, then the sine of $(\theta_2 - \theta_3)$ is small. Thus, each of the preceding derivatives consists of the sum of two small terms, and we might deem them small enough to be negligible.

By contrast, consider the derivatives $\partial P_2 / \partial \theta_3$ and $\partial Q_2 / \partial V_3$. Here, the g's multiply the sine terms, so these terms vanish on both accounts. But this leaves us with the cosine terms multiplied by b's, neither of which are small (since the cosine of a small angle is nearly 1).

$$\begin{aligned}\frac{\partial P_2}{\partial \theta_3} &= |V_2| |V_3| [g_{23} \sin(\theta_2 - \theta_3) - b_{23} \cos(\theta_2 - \theta_3)] \\ \frac{\partial Q_2}{\partial V_3} &= |V_2| [g_{23} \sin(\theta_2 - \theta_3) - b_{23} \cos(\theta_2 - \theta_3)]\end{aligned}$$

Indeed, if we consider the sine terms negligible and the cosine roughly equal to 1, we obtain the following approximations:

$$\begin{aligned}\frac{\partial P_2}{\partial \theta_3} &\approx -|V_2| |V_3| b_{23} \\ \frac{\partial Q_2}{\partial V_3} &\approx -|V_2| b_{23}\end{aligned}$$

Thus, the partial derivatives $\partial P_2 / \partial \theta_3$ and $\partial Q_2 / \partial V_3$ make up the “meat” of the Jacobian matrix. By assuming the small derivatives to all be negligible, we set two of the four partitions of the Jacobian matrix to zero. This sparsity makes it much faster to invert.

As in the dishonest Newton–Raphson method, using a slightly inaccurate Jacobian means we should expect to need more iterations, and we are trading some risk for speed. What is different in the decoupled power flow method is that the simplifications have a physical rationale. If the decoupling was a reasonably good assumption for the network, then we should still be headed in the right direction at each iteration, and should be quite confident to converge to the solution.

12.4.6.3 Fast-Decoupled Power Flow

An even more radical simplification of the Jacobian matrix is possible, called *fast-decoupled power flow*. Here we make a third assumption: that the voltage magnitude profile throughout the system is flat, meaning that all buses are very near the same voltage magnitude (i.e., the nominal system voltage, 1.0 p.u.). We then observe the effect of this assumption, combined with the previous two assumptions about transmission lines and voltage angles, on the Jacobian matrix. By a process of approximation and cancellation of terms, the assumption of a flat voltage profile leads to a much handier version of the Jacobian, including a portion that stays the same during each iteration

²⁹ Recall that $G = R/Z^2$, so that when R approaches zero and there is only reactance ($Z \approx X$), $G \approx 0$ as well.

and therefore saves even more computational effort.³⁰ Again, this should affect only the process of converging on the correct solution, not the solution itself. If the simplifying assumptions were reasonable—in other words, if the simplified derivatives did not lead us in a grossly wrong direction—the computation is vastly sped up.

Note that the power flow solution obtained by the fast-decoupled algorithm will expressly produce a certain profile of voltage angle and magnitude throughout the system that contradicts our literal assumption that these profiles would be flat. Thus, we should think of the flat profiles as merely a procedural crutch along the way to discovering what the true profiles are. The reason we can get away with this apparent conflict is that the iteration process is self-correcting in nature. We can thus incorporate a statement that we know to be false when taken literally (i.e., the voltage profiles are exactly flat) into the directional guidance for our next iteration (the derivatives in the Jacobian matrix), without contradicting the solution at which we ultimately arrive.

Likewise, note the apparent contradiction between the existence of line losses, which can result only from line resistance, and the approximation that the conductances are negligible. Again, the simplifying assumption of ignoring the g's is only a crutch for the process of approaching the correct power flow solution, and the solution itself will be consistent with the actual, nonzero values of conductance and resistance. This solution combined with the explicit resistance values—which, for this purpose, are anything but negligible—then yields the losses for each transmission link.

12.4.6.4 DC Power Flow

The ultimate simplification of a.c. power flow analysis has the misleading name of “DC power flow.” Here we not only assume decoupling and set voltage magnitudes to 1.0 p.u., but we completely ignore reactive power and voltage magnitude to focus exclusively on real power, calculating a solution only in terms of voltage phase angle. The reason it is called DC is that we are dealing with only a single state variable at each node instead of two, as would be the case in a direct-current network, even though the variables still describe an alternating voltage and current.

DC power flow differs from the other shortcuts in that it produces a set of linear equations that can be solved without iteration (just like an actual direct-current network), so it very reliably produces a solution—but not the correct one. The DC power flow solution can only be approximate because the method never computes the actual voltage magnitudes, which are not exactly 1.0 p.u. However, it is by far the fastest way to get a rough handle on a large network. For systems with a reasonable voltage magnitude profile, DC power flow offers an adequate quick overview. This is especially useful if many repetitions of the power flow solution process are required.

One important such application is *contingency analysis* (Section 13.3), where power flow is solved for many different scenarios such as the sudden loss of a transmission line or a generator. This type of analysis is not concerned with getting a highly accurate power flow solution, but with the

³⁰ In the preceding discussion of partial derivatives we have only considered pairs of variables from neighboring buses, that is, the rate of change of real or reactive power at bus i (in our example, 2) with respect to voltage angle or magnitude at one neighboring bus k (in our example, 3). Having chosen that neighboring bus $k = 3$ and taking the derivative, we were able to cheerfully drop the summation sign with all its various k 's, since our rate of change is independent of what happens at all these other buses. However, the Jacobian matrix also contains partial derivatives of power with respect to voltage at the same bus; for example, $\partial P_2 / \partial \theta_2$. These terms, which appear along the diagonal of the matrix, are the ones affected by the assumption of a flat voltage profile. These diagonal terms look different in that they retain the summation over all the other k 's. They succumb, however, to approximation and cancellation of various b 's, leading to vastly simplified expressions. A thorough discussion appears in Arthur Bergen, *Power Systems Analysis* (Englewood Cliffs, NJ: Prentice Hall, 1986).

question of whether a given contingency might result in an egregious violation. If DC power flow raises a flag, that particular scenario can then be studied more carefully.

DC power flow is a true shortcut, which leads not quite exactly to the destination. By analogy, this is like driving directions to a store in a shopping mall that leave you on the wrong side of a parking lot barrier: depending on the nature of your errands, sometimes that's okay.

Besides assuming that voltage magnitudes are 1.0 p.u., DC power flow also assumes that angle differences are small and that line resistances are negligible (i.e., the assumptions for decoupling).³¹ These assumptions produce a linear simplification of the real power flow from Eq. (12.1),

$$P_{ik} \approx \frac{|V_i||V_k|}{x_{ik}} \sin(\theta_k - \theta_i) \approx \frac{1}{x_{ik}}(\theta_i - \theta_k) \quad (12.3)$$

where x_{ik} is the inductive reactance of the lossless line connecting buses i and k , and P_{ik} is positive flowing from bus i to k .

The inverse reactance x can also be written as the admittance or susceptance, b . In fact, the entire admittance matrix \mathbf{Y}_{bus} simplifies into a susceptance matrix if all the entries are purely imaginary, with $y_{ik} = b_{ik}$.³² However, \mathbf{B} is defined as the imaginary components of \mathbf{Y}_{bus} without the row and column corresponding to the slack bus, since we do not solve for the state variables at that bus.

With \mathbf{B} , we can write a concise linear equation for (approximate) real power flow across the entire network, where $\boldsymbol{\theta}$ includes all the bus voltage angles. A negative sign in front of \mathbf{B} accounts for the ordering of the bus angles and admittances in the matrix operations.³³ As before, power is defined as positive when injected into the network (i.e., generated at the bus).

$$\mathbf{P} = -\mathbf{B} \boldsymbol{\theta} \quad \text{or} \quad \boldsymbol{\theta} = -\mathbf{B}^{-1} \mathbf{P} \quad (12.4)$$

Note that the \mathbf{B} matrix only needs to be inverted once and that like the admittance matrix \mathbf{Y}_{bus} it is sparse, so the computational effort should be minimal. The reason we can get away with linearizing the power flow problem in this way is that we are neglecting losses, since there is no resistance in the network as modeled. Consequently, the total amount of power to be generated throughout the system—just the sum of net demand—is known *a priori*, and no iteration is needed.

Out of curiosity, could we reduce the dimensionality of the a.c. power flow problem in the opposite way, by considering only reactive power and voltage magnitude? No, because the total amount of Q that needs to be generated throughout the system depends on reactive I^2X losses on the transmission lines. The analogous simplification would require us to ignore reactance and consider only resistance—but that assumption would seriously disagree with the physical reality of transmission lines, where it is generally true that $X \gg R$.

Example

Consider the three-bus system in Figure 12.16, with two generator buses and one load bus. Bus 1 (generator) is the slack bus. Bus 2 (generator, minus some load) and Bus 3 (load only) are both modeled as PQ buses.³⁴ Our objective is to estimate the power flow on each of the three lines.

³¹ Transformer tap settings are another detail that is ignored.

³² To review, each diagonal element b_{ii} equals the sum of line admittances connected to Bus i , and each off-diagonal element b_{ik} is the negative admittance between Bus i and Bus k .

³³ In an alternative convention, the minus sign in Eq. (12.4) is subsumed within the definition of \mathbf{B} , in which case all the $b_{ik} = -y_{ik}$.

³⁴ Bus 2 might instead be a PV bus, in which case we would be given the bus voltage magnitude instead of reactive power injection. For DC power flow, this wouldn't matter at all.

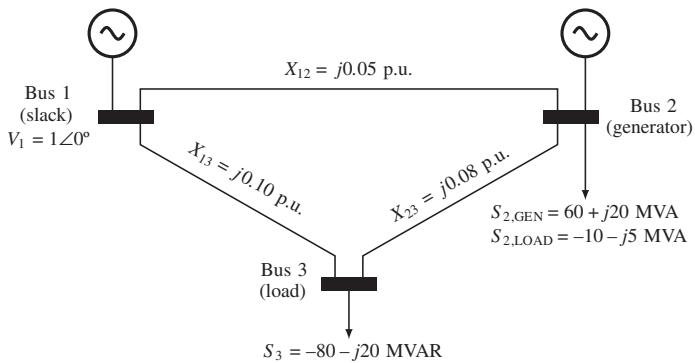


Figure 12.16 Three-bus network.

The power generation and demand are given as follows:

Bus Number	Bus Type	P_G	Q_G	P_D	Q_D
1	Slack	Unknown	Unknown	0	0
2	PQ	60 MW	20 MVAR	10 MW	5 MVAR
3	PQ	0	0	80 MW	20 MVAR

The branch admittances in per-unit are the inverse of the impedances shown in Figure 12.16: $y_{12} = -j20$, $y_{13} = -j10$, and $y_{23} = -j12.5$. The bus admittance matrix is

$$\mathbf{Y}_{\text{bus}} = \begin{bmatrix} -30 & 20 & 10 \\ 20 & -32.5 & 12.5 \\ 10 & 12.5 & -22.5 \end{bmatrix}$$

Since Bus 1 is the slack bus, we have voltage angles at two buses to solve for, θ_2 and θ_3 . Using a base power $S_{\text{base}} = 100 \text{ MVA}$, the net power injections at these buses are $P_2 = 0.5$ and $P_3 = -0.8$ p.u. The information about Q will be ignored entirely.

We are thus solving a linear system with two input and two output variables. Therefore, the \mathbf{B} matrix has to be reduced to the same 2×2 size (in linear algebra terms, it must be of the appropriate rank). This is done by simply deleting the row and column corresponding to the slack bus. In our example, we delete the first row and first column of \mathbf{Y}_{bus} .³⁵

Equation (12.4) then becomes:

$$\begin{bmatrix} \theta_2 \\ \theta_3 \end{bmatrix} = -\mathbf{B}^{-1} \mathbf{P} = \begin{bmatrix} 32.5 & -12.5 \\ -12.5 & 22.5 \end{bmatrix}^{-1} \begin{bmatrix} 0.50 \\ -0.80 \end{bmatrix}$$

The easy-to-find, exact, and guaranteed unique solution (because \mathbf{B} is invertible) to the wrong problem (because it does not quite represent the system under consideration) is

$$\begin{bmatrix} \theta_2 \\ \theta_3 \end{bmatrix} = \begin{bmatrix} 0.002173 \\ -0.03435 \end{bmatrix} \text{ rad} = \begin{bmatrix} 0.124^\circ \\ -1.968^\circ \end{bmatrix}$$

³⁵ Note that the information about impedances connected to the slack bus is not lost, but is still captured in the self-admittances for Buses 2 and 3. Without deleting the slack bus row and column, the matrix would contain redundant information and could not be used to solve a 3×3 system of equations. We would discover this as soon as we tried to invert the matrix.

These results pass an initial sanity test in that the angle at Bus 3 should be moderately negative. The direction of power flow between Buses 1 and 2 is not obvious *a priori*, so a small angle difference is plausible.

From the estimated voltage angles, we can directly calculate power flows on each line, using Eq. (12.3) with angles in radians and $\theta_1 = 0$:

$$P_{12} \approx y_{12}(\theta_1 - \theta_2) = 20 \cdot (-0.00217) = -0.0435 \text{ p.u.}$$

$$P_{13} \approx y_{13}(\theta_1 - \theta_3) = 10 \cdot 0.03435 = 0.3435 \text{ p.u.}$$

$$P_{23} \approx y_{23}(\theta_2 - \theta_3) = 12.5 \cdot 0.03652 = 0.4565 \text{ p.u.}$$

We read this as 4.35 MW flowing from Bus 2 to Bus 1, 34.35 MW flowing from Bus 1 to Bus 3, and 45.65 MW flowing from Bus 2 to Bus 3. The numbers add up correctly for each bus, including the slack bus power injection that we already knew to be 30 MW (since the network is lossless).

12.5 Applications

Power flow analysis is a fundamental and essential tool for operating a power system, as it answers the basic question: What happens to the overall *state* of the system if generation and load change in certain ways, at certain locations? This question may be posed in the context of either day-to-day operations or longer-term planning.

In the short run, a key part of a system operator's responsibility is to approve generation schedules that have been prepared on the basis of some economic considerations—whether by central planning or by competitive bidding—and scrutinize them for technical feasibility. This assessment hinges on power flow studies to predict the system's operating state under a proposed dispatch scenario. If the analysis shows that important constraints such as line loading limits would be violated, the schedule is deemed infeasible and must be changed.

Even with feasible schedules in hand, reality does not always conform to plans, requiring operators to monitor any changes and, if necessary, make adjustments to the system in real time. Power flow analysis is the only comprehensive way to predict the consequences of changes such as increasing or decreasing generation levels, increasing or decreasing loads, or switching transmission links and assessing whether such changes are safe or desirable for the system. Specifically, operators need to know impacts of any actions on voltage levels (are they within proper range?), line flows (are any thermal or stability limits violated?), line losses (are they excessive?), and security (is the operating state too vulnerable to individual equipment failures?). Similarly, power flow analysis is a fundamental tool in the planning context to evaluate changes to generation capacity or the transmission and distribution infrastructure.

12.5.1 Optimal Power Flow

Sometimes it is necessary to compare several hypothetical operating scenarios for the power system to guide operating and planning decisions. Specifically, one often wishes to compare and evaluate different hypothetical dispatches of generation units that could meet a given loading condition. Such an evaluation is performed by an *optimal power flow* (OPF) program, whose objective is to identify the operating configuration or “solution” that best meets a particular set of criteria. These criteria may include the total cost of generation, transmission line losses, and various requirements concerning the system's security, or resilience with respect to disturbances.

An OPF algorithm consists of numerous power flow analysis runs, one for each hypothetical dispatch scenario that could meet the specified load demand without violating any constraints. Clearly, this makes OPF vastly more computation-intensive than just a basic power flow analysis. The output of each individual power flow run, which is a power flow solution in terms of bus voltage magnitudes and angles, is evaluated according to one or more criteria that can be wrapped into a single quantitative metric or *objective function*—for example, the sum of all line losses in megawatts, or the sum of all generating costs in dollars when line losses are included. The OPF program then devises another scenario with different real and reactive power contributions from the various generators and performs the power flow routine on it, then another, and so on until the scenarios do not get any better and one is identified as optimal with respect to the chosen metric. This winning configuration with real and reactive power dispatches constitutes the output of the OPF run. OPF solutions may then provide guidance for on-line operations as well as generation and transmission planning.

Especially for applications in a market environment, where planning and operating decisions may have sensitive economic or political implications for various parties, it is crucial to recognize the inherently subjective nature of OPF. Power flow analysis in and of itself answers a question of physics. By contrast, OPF answers a question about human preferences, coded in terms of quantitative measures. Thus, what is found to constitute an “optimal” operating configuration for the system depends on how the objective function is defined, which may include the assignment of prices, values, or trade-offs among different individual criteria. In short, “optimality” does not derive from a power system’s intrinsic technical properties, but from external considerations.

It is also important to understand that the translation of an OPF solution into actual planning and operating decisions is not clear-cut and has always involved some level of human judgment. For example, the computer program may be too simplistic in its treatment of security constraints to allow for sensible risk trade-offs under dynamically changing conditions, which then calls for some engineering judgment to adapt the OPF recommendation in practice.

At the same time, the computational process is already complex enough that different OPF program packages may not offer identical solutions to the same problem. Therefore, the output of power flow analysis including OPF constitutes advisory information rather than deterministic prescriptions. Indeed, the complexity of the power flow problem underscores the difficulty of managing power systems through static formulas and procedures that would lend themselves to automation, especially if a system is expected to perform near its physical limits.

12.5.2 State Estimation

The set of voltage magnitudes and angles at every network node define the *state* of the system. Power flow analysis is about identifying that state based on a set of input variables, specifically power injections at the various nodes or buses. In practice, however, there may not be sufficiently recent or reliable data available for every bus to determine the system state in a time frame relevant for operational decisions (say, on the order of minutes).

State estimation is the process of combining available information to propose a solution for the system state that best fits the available data. These data typically include direct physical measurements, pseudo-measurements (i.e., values that are known even if not physically measured, such as zero power injection at a node with neither generation nor load connected to it), a network model with impedances and connectivity, Ohm’s law and Kirchhoff’s laws. By corroborating information across multiple sources, state estimation can make up for lacking data at a particular node. Moreover, it recognizes that any given data point could be erroneous, and it can assign different

weights or credibility to different reported values. For example, a measurement may not be properly time-aligned, the instrument may not be calibrated accurately, or there could be noise, delay or intermittency in communication—and some reporting locations might be chronically more suspect than others.

In sum, a state estimation algorithm seeks the state of the system that is most consistent and plausible in view of all the available information. One common approach is the *least squares* or *weighted least squares* fit, used in many other applications of statistical analysis. In essence, it asserts that the solution most consistent with the data is the one that minimizes the collective discrepancies between the empirical measurements and the points proposed by the solution, where discrepancy is defined as the squared difference.

The mathematical problem statement distinguishes the true state vector \mathbf{x} (the set of all bus voltage magnitudes and angles), the function $\mathbf{h}(\mathbf{x})$ of the true state vector that specifies what all the correct measurements should be, and the set of empirical measurements \mathbf{z} that can be expressed as the correct measurements plus an error term $\boldsymbol{\varepsilon}$. Thus, for the i th node in the network, we would write

$$z_i = h_i(\mathbf{x}) + \varepsilon_i$$

Note that each h_i term is a function not only of the state variables x_i at the local bus, but the state \mathbf{x} of the entire network. In this way, $\mathbf{h}(\mathbf{x})$ captures the laws of physics: for example, you cannot reasonably measure a voltage at one bus that is terribly different from its neighbor, or a current that doesn't obey Ohm's law.

If the errors are random and follow a Gaussian distribution,³⁶ the maximum likelihood estimate of the true state based on m measurements is the solution to the least squares problem: find \mathbf{x} that satisfies

$$\min \sum_{i=1}^m \varepsilon_i^2 = \min [\mathbf{z} - \mathbf{h}(\mathbf{x})]^T [\mathbf{z} - \mathbf{h}(\mathbf{x})]$$

(where multiplying by the transpose of the vector is simply the proper way to square it). The weighted least squares formulation includes an *a priori* emphasis (captured by the vector \mathbf{W}) on measurements known to be more accurate,

$$\min \sum_{i=1}^m \varepsilon_i^2 = \min [\mathbf{z} - \mathbf{h}(\mathbf{x})]^T \mathbf{W} [\mathbf{z} - \mathbf{h}(\mathbf{x})]$$

The solution can be found by iterative computation, taking successive guesses at \mathbf{x} and improving the guess with each iteration toward reducing the error.

One important observation is that state estimation requires some redundancy of measurement. If there are n nodes in the network, we need $m > n$ data points for the minimization problem to have a unique solution. A network for which sufficient information is available to produce a unique state estimate is called *observable*.

Many modern control rooms have state estimator applications that issue an updated report every few minutes. Besides providing grid operators the closest thing to a full view of the system in real-time, state estimation can be used to verify input information. Through its iteration process, the state estimation algorithm can identify if a particular data source makes especially egregious

³⁶ In statistics, the bell-shaped Gaussian distribution describes a collection of observations that vary due to random chance.

error contributions. Such an observation can flag bad measurement data as well as bad assumptions about impedance parameters, topology, or other modeling errors. The state estimator may even fail to converge on a solution altogether, in a sign that something is very wrong.

For example, suppose a transmission line tripped offline and is no longer connected, but this event was not reported. Because the connectivity assumed by the state estimator no longer reflects the actual physical relationship among the measurement points, a large error between the empirical and the presumed correct measurements would result, causing a failure of the state estimator to converge. This would in turn trigger an alarm, prompting operators to check for the source of the problem or inconsistency.³⁷

In commercial practice to date, state estimation has been used almost exclusively at the transmission level, although growth in active distributed resources (Section 15.2) motivates the development and adoption of distribution state estimation tools. Distribution systems pose several unique challenges in this context. They are more difficult both to model and to observe, simply because there are so many connecting points for loads, while measurements are scarce (especially in real-time). Also, distribution state estimation requires a three-phase line model that accounts explicitly for each phase and all the mutual impedances, since lines are untransposed (Section 7.2.2) and load imbalance is more pronounced. Yet distribution system models at this level of granularity are notoriously inaccurate, owing to the vast number of variables and the myriad ways in which the real-world condition of the network might change but go unreported. Finally, for the purpose of state estimation, the radial topology of distribution systems—which otherwise facilitates both operation and analysis—is a disadvantage, because it affords less redundancy from Kirchhoff's laws for mutual corroboration of measurements.

12.6 LinDistFlow

Some useful simplifications of power flow analysis can be made for the special case of radial networks, which includes most distribution systems. The *LinDistFlow* equations provide an approximate overview of the relationships between power flow and voltage drop along a radial line.³⁸

The full power flow equations as presented earlier in this chapter specify the relationship between real and reactive power transfer, line impedance, and voltage magnitude and phase angle at either end. As the reader surely appreciates by now, these equations are nonlinear, cumbersome to manipulate, and not readily intuitive. *LinDistFlow*, by contrast, can help build intuition about how voltage phasors relate to power flows, with the aid of a few simplifications. The resulting approximate equations, even though they don't give exact solutions, provide a general sense of how these quantities can be expected to vary. In particular, the effect of capacitors or distributed generation on voltage drop can be helpfully approximated with *LinDistFlow*.

Again, this analysis applies only to radial systems, or radial regions within a network. Specifically, this means that any circuit branch to be analyzed is a single branch connecting two nodes (labeled

³⁷ When the above scenario occurred in Ohio on August 14, 2003, this alarm happened to be disabled and operators remained unaware of a crucial transmission line trip in their area—one of several factors leading to the Northeast Blackout. See *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, U.S.-Canada Power System Outage Task Force, 2004.

³⁸ The terminology and derivation originates from a seminal paper by Mesut Baran and Felix Wu, "Network Reconfiguration in Distribution Systems for Loss Reduction and Load Balancing," *IEEE Transactions on Power Delivery* 4(2), 1401–1407, 1989. The presentation in this section is based on work by Roel Dobbe, Michael Sankur, and Dan Arnold.

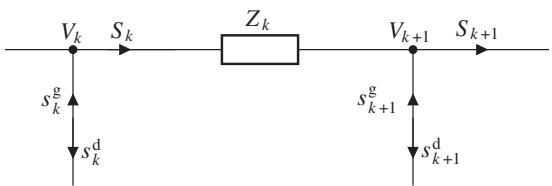


Figure 12.17 Radial distribution branch for the derivation of LinDistFlow equations.

in Figure 12.17 as k and $k + 1$), with no other parallel circuit branch or “back way” connecting those two points. Our objective is to write an expression that relates complex power \mathbf{S}_k flowing from node k to node $k + 1$ across the line impedance $\mathbf{Z}_k = R + jX$, and the voltage phasors \mathbf{V}_k and \mathbf{V}_{k+1} at the two nodes.³⁹ For generality, we also include power demand and generation s^d and s^g that result in some net power injection at each node, but these will simply add to the power entering and leaving each node to other directions and be subsumed within \mathbf{S}_k or \mathbf{S}_{k+1} .⁴⁰ What matters here is that we can isolate the power flow \mathbf{S}_k on the single branch connecting any two nodes in question. We will simply let $k = 1$ for the following discussion.

Let us present the LinDistFlow equations first and appreciate the insights they convey, and then return to their derivation. The linearized approximation for voltage magnitudes⁴¹ V_1 and V_2 at adjacent nodes in a radial system is

$$V_1^2 - V_2^2 \approx 2(RP + XQ) \quad (12.5)$$

Note that Eq. (12.5) describes a linear relationship, even though the voltage magnitudes appear squared, because the square terms are completely isolated and we can simply redefine a variable such as $E_1 = V_1^2$.⁴² The main assumption for this approximation is that power losses are negligible. This will be closer to true when impedances and power flows are small.

A key point illustrated by Eq. (12.5) has to do with the relative importance of real and reactive power for voltage magnitude drop, depending on the relative dominance of resistance or reactance on the line. Recall the general rule that voltage drop is more sensitive to reactive power than real power transfer. In truth, this rule applies only to lines that are mostly inductive, and especially when $X \gg R$.⁴³ That condition happens to hold true for almost all transmission lines and, to a lesser degree, for many distribution lines.

Equation (12.5) also tells us, though, that in case a line’s resistance is comparable to its reactance, then real power flow will contribute appreciably to the voltage drop. The relative impact of P and Q could even be reversed if, hypothetically, a line’s resistance exceeded its reactance. In practice, the impact of distributed generation on feeder voltage depends significantly on the line’s X/R ratio. In any case, because of the plus sign in Eq. (12.5), both real and reactive power flow from node 1 to 2 are associated with the voltage magnitude decreasing from 1 to 2.

³⁹ It turns out that it doesn’t matter if we label the power flow S_k or S_{k+1} . The reason is twofold: In the voltage magnitude equation, we will assume that there are no losses. Therefore, power out of node 1 equals power into node 2, and the subscript assignment is a matter of stylistic preference. In the angle equation, we have a difference of phase angles, which will be the same regardless of whether we take node 1 or 2 as the reference.

⁴⁰ The generality is important in case we want to string together multiple branches and use consistent labels to account for all nodes in relation to each other. This will work as long as the system is radial.

⁴¹ In this section, we drop the absolute value signs and use boldface notation for phasors.

⁴² Solving for V^2 still leaves two mathematically possible solutions for V , but one is negative and therefore physically implausible.

⁴³ Note that this condition and conclusion are perfectly consistent with our earlier analysis of the full nonlinear power flow equations and the Jacobian matrix. LinDistFlow just makes the relationship much easier to see.

The corresponding relationship for voltage phase angles can be shown at any of three successive levels of simplification, the last two of which qualify as linear equations:

$$\sin(\delta_1 - \delta_2) = \frac{1}{V_1 V_2} (XP - RQ) \quad (12.6)$$

$$\delta_1 - \delta_2 \approx \frac{1}{V_1 V_2} (XP - RQ) \quad (12.7)$$

$$\delta_1 - \delta_2 \approx (XP - RQ) \quad (12.8)$$

The first form, Eq. (12.6), is exact and requires no specific assumptions other than the radial topology. For the case where $X \gg R$, the RQ term becomes negligible, and the expression reduces to the common approximation for real power flow versus angle (as in Section 7.3.2).

Because voltage phase angle separations in distribution systems tend to be small, we can get the linearized equation (12.7) by making the small-angle approximation $\sin \theta \approx \theta$ (where θ is expressed in radians, not degrees), which is usually excellent.⁴⁴ Getting rid of the trigonometric function is a huge advantage, especially if we want to use these variables in some linear algebra formulation.

A further simplification, shown in Eq. (12.8), assumes that the voltage magnitudes are both equal to 1.0 p.u. (Section 8.7), and the fraction with V_1 and V_2 simply goes away. This allows us to make a statement about voltage phase angles without any information about the magnitudes, which can be very helpful.

This “well-behaved voltage” approximation will tend to introduce a greater numerical error than the small-angle approximation, but it won’t qualitatively affect the relationship of variables to each other. Typical voltage drops in distribution systems are on the order of single-digit percent, as the general operational standard calls for maintaining voltage magnitudes in the range of 0.95 to 1.05 p.u. (95% to 105% of nominal). In the worst case that still meets this standard, we would have about a 10% error:

$$\frac{1}{V_1 V_2} = \frac{1}{0.95 \cdot 0.95} = 1.108 \quad \text{or} \quad \frac{1}{V_1 V_2} = \frac{1}{1.05 \cdot 1.05} = 0.907 \text{ p.u.}$$

The relationships described by the pair of LinDistFlow equations for voltage magnitude and angle are especially useful in the context of a modern system with distributed generation, where both real and reactive power might flow in either direction. For example, there may be very little difference in voltage magnitude due to opposing effects of P and Q flow, while the current is greater than expected. If current measurements are not available, complementing voltage magnitude information with voltage angle measurements would provide conclusive information about whether P and Q are both small or acting in opposition.

The associations of voltage magnitude and angle profiles with positive, zero, or negative real and reactive power flows, and the dependence of this association on the impedance of the line, are illustrated in Figure 12.18 for two special cases, with $X \gg R$ ($R \approx 0$) and $R \approx X$. Figure 12.18a illustrates the familiar result from the transmission context: namely, that real power flows from greater to smaller voltage phase angle, and reactive power flows from greater to smaller voltage magnitude. In distribution systems, however, line resistance may be much more significant, and $X \gg R$ may be a poor approximation. Figure 12.18b illustrates the special case where resistance and reactance are equal. The result is a clockwise rotation of the iso- V and iso- δ lines by 45°.

One intuitive insight conveyed by this qualitative analysis is the crucial importance of the X/R ratio for characterizing distribution lines, and for predicting the impacts of distributed generation

⁴⁴ For example, with a phase angle difference of 1°, we have $\sin 1^\circ = \sin 0.017453 \text{ rad} = 0.017452$ and the approximate equality is good to four significant figures.

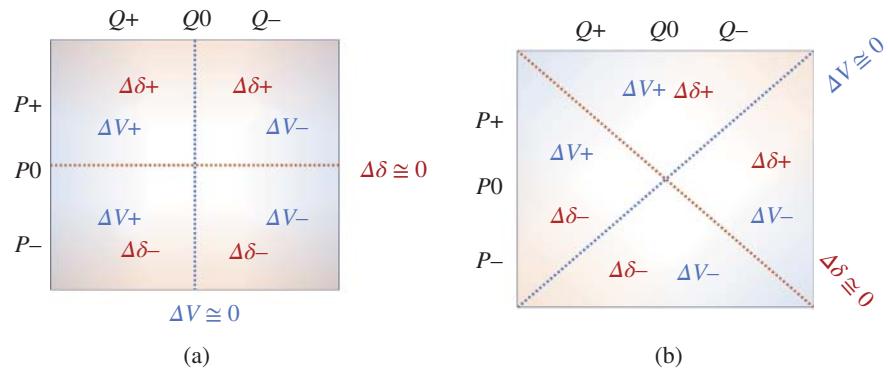


Figure 12.18 Qualitative association of voltage drop $\Delta V = V_1 - V_2$ and angle difference $\Delta\delta = \delta_1 - \delta_2$ with the direction of real and reactive power flow, based on Eqs. (12.5) and (12.8). $P+$ and $Q+$ indicate positive power flowing from Node 1 to 2, in the sense of S_k in Figure 12.17. $\Delta V+$ indicates that $|V_1| - |V_2| > 0$, and $\Delta\delta+$ that $\delta_1 - \delta_2 > 0$. (a) When $X \gg R$, the sign of $\Delta\delta$ is aligned with real and ΔV with reactive power flow. (b) When $X \approx R$, the relationship rotates by 45°.

(Section 15.2.2). Another observation is that for managing voltage on radial distribution feeders, real and reactive power control can play complementary roles. This motivates the dual option for both volt-VAR and watt-VAR droop curves for inverter control (Section 14.4.3).

12.6.1 Derivation

To derive LinDistFlow equations, we first write down Kirchhoff's voltage law (KVL), stating that one nodal voltage is given by the sum of the other, plus the voltage drop between them:

$$\mathbf{V}_1 = \mathbf{V}_2 + \mathbf{I}Z$$

It does not matter which node we label as 1 or 2, as long as the current is defined as positive in the direction consistent with KVL. We use boldface notation here to remember that these are all complex phasor quantities.

Our objective is to obtain separate, real expressions for voltage magnitudes and angles, respectively. The general trick is to multiply complex expressions by their complex conjugate.

First, for the voltage magnitude equation, we multiply both sides of the equation by \mathbf{V}_1^* :

$$\begin{aligned}\mathbf{V}_1 \mathbf{V}_1^* &= (\mathbf{V}_2 + \mathbf{I}Z)(\mathbf{V}_2^* + \mathbf{I}^*Z^*) \\ &= \mathbf{V}_2 \mathbf{V}_2^* + \mathbf{V}_2 \mathbf{I}^* \mathbf{Z}^* + \mathbf{I}Z \mathbf{V}_2^* + \mathbf{I}Z \mathbf{I}^* \mathbf{Z}^*\end{aligned}$$

Initially, the result of cross-multiplying these terms looks confusing. But realizing that the product of a complex number and its complex conjugate is just the magnitude squared, and employing the identity $ab^* + a^*b = 2 \operatorname{Re}\{ab^*\}$, we get

$$V_1^2 = V_2^2 + 2 \operatorname{Re}\{\mathbf{V}_2 \mathbf{I}^* \mathbf{Z}^*\} + I^2 Z^2$$

The last term represents a measure of line losses multiplied (again) by impedance. We can reasonably expect this term to be small (since the line impedance should be a small number, especially when squared) and will choose to neglect it; this is the essence of the linearization. Consequently, the equality is only approximate from here on.

Rearranging terms and substituting complex power $\mathbf{S} = \mathbf{I}^* \mathbf{V}_2$ we get

$$V_1^2 - V_2^2 \approx 2 \operatorname{Re}\{\mathbf{S} \mathbf{Z}^*\}$$

which expands into

$$V_1^2 - V_2^2 \approx 2 \operatorname{Re}\{(P + jQ)(R - jX)\} = 2(RP + XQ)$$

where taking the real part retains half the pairings from the cross-multiplication to yield Eq. (12.5).

For the derivation of the angle equation, we again start with KVL, but now take the complex conjugate of both sides and multiply by \mathbf{V}_2 , to obtain

$$\mathbf{V}_1^* \mathbf{V}_2 = \mathbf{V}_2^* \mathbf{V}_2 + \mathbf{I}^* \mathbf{Z}^* \mathbf{V}_2$$

Notice that the product $\mathbf{V}_1^* \mathbf{V}_2$ will give us the difference between the two voltage phase angles.

We again substitute complex power $\mathbf{S} = \mathbf{I}^* \mathbf{V}_2$ and this time take only the imaginary part of both sides of the equation, which produces the sine on the left hand side and retains the opposite pairings of cross-multiplied terms on the right,

$$\begin{aligned} V_1 V_2 \sin(\delta_1 - \delta_2) &= \operatorname{Im}\{\mathbf{S} \mathbf{Z}^*\} \\ &= \operatorname{Im}\{(P + jQ)(R - jX)\} = RQ - XP \end{aligned}$$

yielding Eq. (12.6).

Problems and Questions

- 12.1** Repeat the iterative Newton–Raphson solution of the two-bus power flow example with a line impedance of $Z = j0.05$ p.u. How do you expect the solution for V_2 and the losses will be different?
- 12.2** Repeat the iterative Newton–Raphson solution of the two-bus power flow example with a (physically unrealistic!) purely resistive line with $Z = 0.1$ p.u. How do you expect the solution for V_2 and the losses will be different?
- 12.3** Repeat the iterative Newton–Raphson solution of the two-bus power flow example with a starting guess of $\mathbf{x} = [0, 0.25]$ p.u., and comment.
- 12.4** For the three-bus example in Figure 12.19, use the DC power flow approximation to estimate real power flow on each of the three lines.

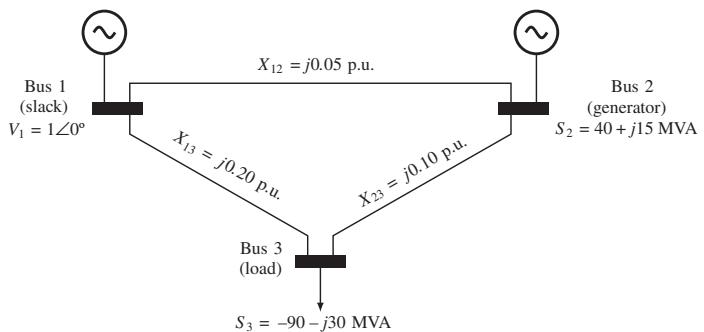


Figure 12.19 Three-bus example for Problems 12.4 and 12.5.

- 12.5** Find the full power flow solution to the power flow example in Figure 12.19 using the dishonest Newton–Raphson method until the mismatch is within 10^{-3} , and compare to the answer from the previous problem. Besides giving no information about reactive power injection and voltage magnitudes, how wrong was the DC Power Flow solution on the angles θ_2 and θ_3 ?
- 12.6** Consider a single-phase distribution line segment with an impedance Z between nodes 1 and 2. You are given the information that $V_1 = 1.000\angle 0^\circ$ p.u. and that $P_2 = 0.40$ p.u. and $Q_2 = 0.20$ p.u. Estimate V_2 using the LinDistFlow equations (express δ in degrees), for the following cases:
- The line impedance is $Z = 0.050 + j 0.050$ p.u.
 - The line impedance is purely resistive, with $R = 0.0707$ p.u. (Note: this is a physically unrealistic scenario.)
 - The line impedance is purely inductive, with $Z = jX = j 0.0707$ p.u.
 - Qualitatively, explain how these cases are different.
- 12.7** Suppose you wish to minimize the voltage magnitude drop across the line segment in the previous problem by adding a capacitor at Node 2. You choose a capacitor that injects $Q_{\text{CAP}} = 0.20$ p.u. to exactly offset the reactive power flow into Node 2. What is the effect of the capacitor on the voltage magnitude $|V_2|$ in each of the above cases of line impedance?
- 12.8** Do you have reason to believe that the linear approximation is good for the previous problem? Discuss.

13

Limits

This chapter combines several topics from very different branches of engineering analysis. The common theme is the envelope or constraints within which the system can be operated, and our confidence that requirements are met. What are the limits? Will the lights stay on, and how can we be sure? The main concepts presented here are *adequacy*, *reliability*, *security*, and *stability*.

13.1 Adequacy

Resource adequacy is concerned with the question, will there be enough electric generation supply to meet demand? Adequacy has both a real-time operations and a longer-term planning component. On any given day, one wants to ensure that there are enough back-up resources to draw upon in case demand turns out to be greater than forecast, or in case of a contingency where some generation or transmission capacity suddenly becomes unavailable. That means some amount of generation in excess of the highest anticipated load needs to be committed and ready to go in case it is needed. Looking toward future load growth and recognizing the time it takes to permit and build new generation facilities, guaranteeing resource adequacy also means planning years in advance to ensure that resources will be ready when called for.

Historically, the analysis of resource adequacy has emphasized the generation aspect, especially at the system level. In part, this is because transmission systems were designed with enough excess capacity to merit the assumption that generated power could always be delivered, anywhere. As transmission systems are being more fully utilized due to a combination of economic pressures, demand growth, interconnections between territories, and difficulties in siting new lines, transmission is becoming an increasingly important constraint. For example, there may be plenty of generation resource in one area, but no realistic way to deliver the power to another area where it is needed.

There is no single correct answer for how much extra generation resource is definitely enough, just like there is no single correct answer for how much insurance coverage one should buy. Resource adequacy requirements reflect a collective policy decision about managing risk. Different jurisdictions and regulatory agencies employ different standards of resource adequacy that they require of electricity providers or *load serving entities*. For example, a utility might be asked to demonstrate to their balancing authority or their public utilities commission that they have contracts in place to procure some number of megawatts to meet their resource adequacy obligations for the coming peak demand season, to ensure that they will be able to serve all their customers. Setting an explicit bar for such obligations not only prevents cutting corners; it also

serves as a benchmark for judging whether reasonable efforts have been made, in the event that the system does fall short.

A standard metric for resource adequacy is the *reserve margin* of available generation resources, expressed as a fraction or percentage of total load. Formally, the *planning reserve margin* (PRM) is defined as

$$\text{PRM} = \frac{\text{firm capacity} - \text{peak demand}}{\text{peak demand}}$$

where *firm capacity* refers to the projected generation resource in megawatts.

Historically in the United States, planning reserve margins of 20% have been considered standard, with some as high as 25%; more recently, margins around 15% are common. One weakness of the reserve margin approach is that it does not fully account for the characteristics of specific generation units, notably their various failure rates, which may differ significantly. With better information, it is possible to operate a system at the same level of confidence with a tighter margin. Conversely, even a seemingly conservative margin might prove to be inadequate, if incorrect assumptions were made in the risk analysis. For example, errors could be due to load forecasting, generation forecasting, accounting for outages, crediting available capacity, or resources not performing in real-time as promised.

During a supply shortage, as the reserve margin shrinks below certain thresholds, the system operator may call for various levels of emergency measures, including the most expensive supply and any available demand response, along with voluntary conservation efforts. The goal is to avoid the need to forcibly disconnect or *shed* load. As a last resort, shedding load in *rotating outage blocks* of customers aims to spread the burden of power outages equitably among customers while preventing the entire system from collapsing with underfrequency. Once the power to customers goes out, system reliability is measurably impacted.

13.2 Reliability

Reliability generally describes the continuity of electric service to customers, which depends both on the availability of sufficient generation resources to meet demand and on the ability of the transmission and distribution system to deliver the power. Most power outages in the United States are caused by local disruptions to the distribution infrastructure, and not by any general shortage of supply. While some redundancy is built into transmission systems (analyzed in terms of *security* below), *radial* distribution systems (see Section 7.1.5) are particularly vulnerable to disruptions because there is only one path for power to flow. However, these local failure probabilities vary greatly (e.g., depending on topography and weather) and are difficult to capture in predictive analysis.

13.2.1 Measures of Reliability

A standard target metric for reliability is the *loss-of-load probability* (LOLP), which states the probability that during any given time interval, the systemwide generation resources will fall short of demand. Classically, this probability is derived from the failure probabilities of the individual generators (i.e., the chance of that generator being unavailable) by summing up the probabilities of all the possible combinations in which the total capacity is less than the anticipated load. The LOLP may be considered on a daily basis (looking at the peak load for that day) or for each individual hour.

A closely related measure is the *loss-of-load expectation* (LOLE), in which the probability of loss-of-load for each day is summed up over a time period and expressed as an inverse, to state that we should expect one loss-of-load event during this period. The smaller the LOLP, the longer on average we will go until an outage happens. For example, if the LOLP is 0.00274 (1/3650) every day, this corresponds to a LOLE of one day in 10 years. In other words, the systemwide generation capacity is expected to fall short of demand, presumably at the peak demand hour of that day, once every 10 years. This latter figure has traditionally served as a benchmark value for reliability—the “one-day-in-ten-years criterion”—throughout the utility industry in the United States.

Note that the LOLP does not capture the effects of contingency events throughout the transmission and local distribution system. Also, the LOLP and LOLE say nothing about the duration of an outage; in particular, one day in 10 years does not mean the load will be interrupted for all 24 hours of that day.

The *expected unserved energy* (EUE) can be calculated by combining the probability of loss-of-load with the actual megawatt (MW) amount of load that would be in excess of total generating capacity. This process assumes that the excess load would be *shed*, or involuntarily disconnected so as to retain system integrity and continue to serve the remaining load.

As measures of systemwide properties, the above terms refer to a broad footprint or jurisdiction, as defined by a utility’s service territory or a balancing authority.¹ Again, they consider only outages due to generation shortfall, not local disturbances in the transmission and distribution system.

In truth, however, failures of the power delivery infrastructure are by far the most common cause of service interruptions. For this reason, service reliability varies regionally, depending in large part on topography and climate as well as population density. In the mountains, for example, distribution lines are much more prone to storm damage, and it will take service crews longer to reach and repair them. Moreover, where only a small number of customers are affected by a damaged piece of equipment, its repair will be lower on the utility’s list of priorities, especially after a major event when line crews are working around the clock to restore service. In major urban areas, by contrast, distribution systems are often placed underground or designed with more redundancy to improve reliability, and the additional costs are justified by the high load density along with the presence of sensitive or critical loads.² The actual service reliability for specific customers within a power system is therefore a highly variable quantity that depends on many factors.

The actual service reliability experienced can be quantified in terms of how often service to certain loads is interrupted (an *outage* occurs), and how long the interruption lasts: *outage frequency* and *outage duration*. The product of outage frequency and average duration gives the total outage time. Since the most typical service interruptions are those associated with events in the distribution system, some of which can be very brief (such as the operation of a reclosing circuit breaker to clear a fault, Section 7.5), outage frequency may be computed so as to include only interruptions lasting longer than a specified time. In fact, there is a continuum from *power quality* to reliability; what constitutes a nuisance outage also depends on the sensitivity of loads to brief changes in voltage (Section 5.1).

¹ For the purpose of this analysis, the territory can be defined at any scale since generation resources outside the territory are simply considered as imports. The imports, like native generators, have some probability of being available at any given time.

² Sensitive or critical loads might include, for example, street lights in busy intersections, elevators in high-rise buildings, commercial customers where loss of power may have a big economic impact, police and fire stations, hospitals, public drinking water supply, and government or corporate offices with perceived political significance. While *critical load* is an official designation, distribution operators may also have an informal understanding of which loads take high priority.

Reliability statistics can be gathered and expressed on a per-customer or a systemwide basis. Formal metrics include the *Customer Average Interruption Frequency Index* (CAIFI), *Customer Average Interruption Duration Index* (CAIDI), *System Average Interruption Frequency Index* (SAIFI), and *System Average Interruption Duration Index* (SAIDI).

13.2.2 Valuing Reliability

Generally speaking, the sudden loss of electric service entails some combination of nuisance and/or loss of economic productivity for different customers. For many people, it is simply an inconvenience to have the lights go out, or perhaps a financial burden to have a freezer full of food spoil, or to close their small business for the day. In other cases, there is a direct threat to human health and safety. Classic situations where electricity may be absolutely vital include winter heating, air conditioning on extremely hot days, medical equipment, and traffic signals. Many critical infrastructure facilities, where power interruptions have immediate life-threatening consequences, are equipped with *uninterruptible power supplies* (UPS) fed by their own backup generators.³ Nevertheless, since backup generation may not last indefinitely (e.g., if diesel generators run out of fuel), service restoration to these critical loads remains a top priority.

Although the situations where human lives depend on electric service constitute only a small fraction of electricity uses, they motivate the general importance ascribed to grid reliability. Considering all electric demand to be vitally important makes sense in that an outage does not usually discriminate between the more and the less important loads. For example, plugged in somewhere among the lights, televisions and refrigerators on a city block might be a dialysis machine. Unable to isolate the most vital loads to serve exclusively, power system operators often find themselves responsible to maintain or restore service to all customers with similar urgency.

In turn, this customary high standard of reliability has led industrialized societies to become increasingly dependent on uninterrupted electric service without considering this a risk or vulnerability. For example, gas-fired furnaces with electronic ignition make their owners dependent on electricity to stay warm, even though it isn't needed as the primary energy source.

As mentioned earlier, the one-day-in-ten-years criterion has served as a benchmark for service reliability in the U.S. electric utility industry for many years. From a market perspective, though, the concept has been criticized for its arbitrariness and overgeneralization. An influential study in 1972 charged that much more was being spent on overdesigning equipment than could rationally be justified through the value of that increased reliability to consumers, and that, in this sense, utilities were "gold-plating" their assets.⁴

Historically, utilities' pursuit of very high levels of service reliability has had several reasons. One reason is their legal obligation to serve, as demanded by the regulatory contract that grants them a territorial monopoly in return for the promise to serve all customers without discrimination and to the best of their ability. Associated with this obligation has been a ratemaking process that allowed utilities to recover a wide range of reliability-related investments and expenses through the rates they charge customers, in which convincing public utility commissions of the "prudence" of these investments has not traditionally been very difficult.

The commitment of utilities and regulators alike to investments in system upgrades must also be viewed in the context of electric demand growth, which in the United States was very high during

³ The defining characteristic of a UPS is that it transitions smoothly from main to backup source, without even a momentary interruption to the load.

⁴ Michael Telson, "The Economies of Alternative Levels of Reliability for Electric Power Generating Systems," *Bell Journal of Economics* 6(2), 679–694, 1975.

the period following World War II until the energy crises of the 1970s, and which subsequently tended to be overestimated by analysts who projected continued exponential growth at similar rates. The experience of continuous growth in combination with the fear of energy shortages explains the historical readiness to invest large sums of money in added generation capacity, as well as the focus on reliability measures that emphasize generation shortfalls over other causes of service interruptions.

Finally, commitment to service reliability can also be understood in terms of a culture of workers who see themselves in the role of providing a vital public service, and who have long cultivated a sense of ownership of their vertically integrated system in which they take considerable personal pride. The implications of changing this cultural variable in the competitive market environment are still far from clear.

As of the early 2020s, electric reliability statistics in the United States have indicated some decline, with economic factors as well as extreme weather playing a role. Climate change presents a complex challenge, not only due to increasing variance of temperature and precipitation and greater frequency of extreme events, but also in that the seasonal timing and geography of challenging weather simply differs from what planners were expecting.⁵ Increasing uncertainty is a problem in and of itself. Further, the interaction of meteorological and ecological effects—in particular, causing unprecedented wildfires—can have a highly nonlinear impact on the electric grid and on local or regional reliability.

In the face of these challenges, there is no simple answer for what is the correct amount of money to spend on reinforcing the electric power infrastructure, or how best to do it. Future strategies may include very different approaches, such as placing transmission and distribution lines underground, or developing local microgrids that can be operated as power islands.

An important distinction in this context is between reliability and *resilience*. While reliability focuses on the occurrence of power interruptions, resilience refers to the ability of the system to recover from disturbances and to minimize harm. Thus, strategies to increase resilience might include expanded or faster switching options for reconfiguring distribution system topology, added resources and storage for backup power, or more refined distinctions between critical and non-critical loads.

From an economic perspective, one would ideally determine the optimal level of investment based on the value of reliability, or the cost of being without power. A major difficulty is that these costs vary so widely across society, and across specific uses of electricity. One approach is to quantify customers' *willingness to pay*, which implies disaggregating various aspects of service quality and distinguishing among customer groups with different preferences. Analytically, the problem is to determine what level of reliability is "optimal" for a given type of customer, in that the amount of money spent on providing this level of service is commensurate with the amount this customer would be willing to pay for it, given the option. Such a determination requires a mechanism by which customers can express their preferences. Electricity markets aim to achieve this goal by providing customers with more and increasingly differentiated choices.

The most common approach is to offer rate discounts in exchange for an agreement to disconnect loads whenever the utility deems necessary, up to some maximum number of instances or duration per year. Actually providing different levels of service reliability for various sets of customers then requires a technical mechanism to discriminate among them, or selectively interrupt their

⁵ For example, the 2021 winter storm Uri caused massive power outages in Texas not just because it was very cold, but because the temperatures and associated vulnerabilities came as a *surprise* to those who designed and managed the infrastructure.

service, in order to verify or enforce compliance with the interruptible load agreement.⁶ To date, interruptible service contracts are relatively common for large commercial and industrial customers, but not at the residential level.

There exists a literature on the valuation of electric-service reliability that attempts to identify and distinguish how much service reliability is worth to different types of customers, or to specifically estimate the costs these customers incur as a result of outages. The simplest approach assumes a linear relationship between outage cost and duration. Here, outage cost is expressed in terms of dollars per kilowatt-hour (kWh) lost, where the lost kWh are those that would have been demanded over the course of the outage period. Such a cost might be derived, for example, from the lost revenues of a business during that time. A more refined approach estimates cost components of both outage frequency and outage duration. In the absence of real choices, though, these estimates suffer the same uncertainties as any contingent valuation data that are based on people's responses in surveys, which may differ from the preferences they would reveal in real situations.

More fundamentally, such economic valuations cannot capture the cost of lives lost due to power outages. The absolute need to protect human health and safety stands in contrast with the reality that a 100% reliable electric grid can never be guaranteed, and that there will be diminishing returns on investments to attain more nines (as in 99.99%) of reliability for the system as a whole. Consequently, local solutions to protect people from harm during power outages may become increasingly important, as climate and weather-related disruptions increase in frequency and severity. Some of the enabling technologies include distributed generation (Section 15.2), storage (Section 15.3), and more broadly, information and control technology to help coordinate diverse small-scale resources.

13.3 Security

Security refers to the width of the operating envelope, or set of available operating configurations that will result in a successful outcome, meaning that no load is interrupted and no equipment is damaged. In other words, security describes how many things can go wrong before service is actually compromised. A system in a secure operating state can sustain one or several *contingencies*, such as a transmission line going down or a generator unexpectedly going off-line, and continue to function without interruption, by transitioning into a new configuration in which the burden is shifted to other equipment (the load on other lines and/or generators is suddenly increased).

When a power system with a given physical infrastructure faces an increasing load, the number of alternative operating configurations diminishes, and the system becomes increasingly vulnerable to disturbances. In the extreme case, with all generators fully loaded and with all options to purchase power from outside the system already exhausted, then if one generator fails, some service will inevitably be interrupted.

To avoid this type of situation, utilities have traditionally retained a reserve margin of generation. Increasing interconnections between service territories over the past decades have enabled the confident operation with lower reserve margins than the traditional 20%, since reserves are in effect “pooled” among utilities. At the same time, this approach to providing reliability through scale implies an increased dependence on transmission links, as well as an increasing vulnerability to disturbances far away.

⁶ It is not unheard of for customers to gladly receive a discount, but fail to switch off their load when the call comes at an inconvenient time.

Analogous to generation reserve, system security relies on a “reserve” of transmission capacity, or alternate routes for power to flow in case one line suddenly goes out of service. The analysis of such scenarios is called *contingency analysis*. A standard criterion in contingency analysis is the *N-1 criterion* (for “normal minus one”), which holds that the system must remain functional after one contingency, such as the loss of a major line. For even greater security, an *N-2 criterion* may be applied, in which case the system must be able to withstand two separate contingencies simultaneously. A related but distinct criterion is *N-1-1*, where the two contingencies are assumed to occur sequentially rather than simultaneously.

Note that security criteria are possible to meet only in a system with some redundancy. A strictly radial distribution system (Section 7.1.5) cannot be *N-1* secure, because the failure of a single line or transformer will interrupt the only link to all downstream sections.

Security criteria find expression in the form of *line flow limits*, which state the amount of current or power transfer permissible on each transmission link. The implication is that as long as the currents on all the lines are within their limits, the resulting operating state does not violate any constraints even if one line is lost. This means that immediately after loss of the line, loading on the other lines and transformers will not exceed their ratings, and all voltages can be held within the permissible range. For this purpose, *emergency ratings* may be defined that allow exceeding the normal thermal rating temporarily (say, for 15 minutes) while giving operators time to take some remedial action. Stability limits may also apply; see Section 7.3 on line loading constraints.

Figure 13.1 illustrates a toy example where *N-1* security considerations would limit the permissible power flow on a pair of transmission lines. In this case, each of the two lines that make up the transmission path has a thermal rating of 100 MW, and an emergency rating of 120 MW. For the network to be secure, either line must be able to absorb the power flow due to the possible failure of the other line, without exceeding its emergency rating. This security constraint limits the power flow to 60 MW on each line.

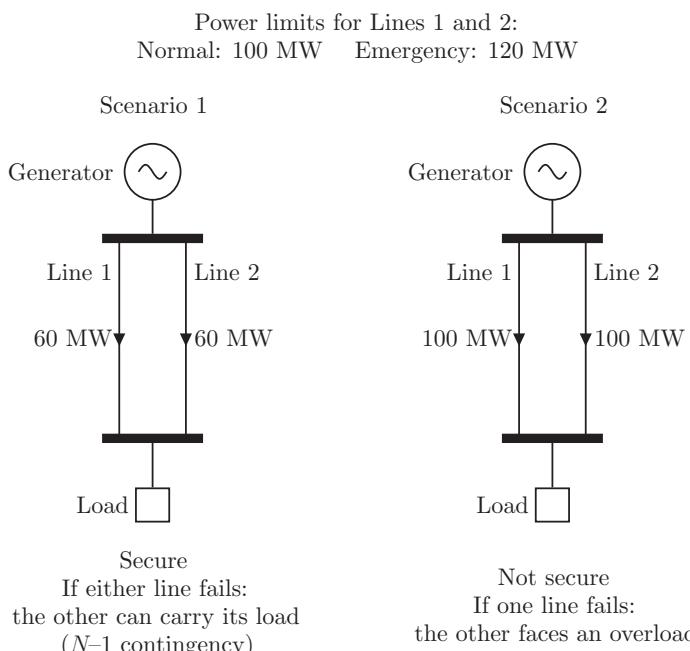


Figure 13.1 Toy example illustrating *N-1* security.

Real-world contingency analysis is much more complicated, not only because there may be many contingencies to consider, but because the effects of each are not obvious. For example, the loss of a single major transmission line or generator could affect line flows in many other parts of the network. This computation-intensive analysis involves running many power flow scenarios (Chapter 12) for a set of load conditions, including peak loads, each time with a different contingency or combination of contingencies, and check that all constraints are still met. Typically, the contingencies are selected from a list of worst-case yet credible possibilities prepared by operators based on experience. It is not always obvious how wide a net to cast when studying contingencies, since distant events in a neighboring jurisdiction or failures of presumably minor components can sometimes have a surprising impact.⁷

The results of contingency analysis are used to set limits for secure operation in *security-constrained dispatch* (Section 16.1.3). They can also help identify priorities for capacity reinforcements in transmission planning.

In the general form described here, contingency analysis is a *steady-state* analysis, meaning that it considers the system operating state before and after the contingency, but not during the event and the transition into the new state. However, that transition itself may pose potential problems; this is assessed in a *dynamic* analysis. Here, contingencies are selected from a shorter list of more serious “dynamic contingencies,” and the system is analyzed for transient and voltage stability during the transition.

13.4 Stability

13.4.1 Overview

In general, *stability* describes the tendency of a system to return to some equilibrium state when perturbed. For the electric grid, this means the tendency to maintain a synchronous and balanced operating state. That operating state is defined by voltage magnitudes and angles at different points throughout the network.

Most often, the term “stability” in a.c. power systems implicitly refers to *angle stability*, which means that all the system’s generators and loads remain locked “in step” at a given frequency: in other words, the generator angles tend to return to an equilibrium position when disturbed. When voltage magnitude is of primary concern, we speak of *voltage stability*. In the transmission context, voltage stability is based on the inherent limitations of a.c. power transfer over long distances.

The two dimensions of voltage magnitude and angle are in fact physically interrelated, but have different sensitivities to control actions. They are often considered separately, both to make the math tractable and to draw straightforward actionable conclusions. Thus, different categories of stability analysis focus on different sets of variables and actuators such as generator torque, transformer load tap changers (Section 7.4), or inverter control (Section 14.4.3).⁸ A separate concept is *frequency stability*, which relates to the systemwide balancing of instantaneous power as described in Chapter 11 and is not further discussed here.

⁷ For an instructive case study, see the FERC and NERC Staff Report, *Arizona-Southern California Outage Event on September 8, 2011—Causes and Recommendations*. Available: https://www.nerc.com/pa/rrm/ea/September%202011%20Southwest%20Blackout%20Event%20Document%20L/AZOutage_Report_01MAY12.pdf (accessed October 2023).

⁸ We are barely scratching the surface here. An authoritative reference on the subject is P. Kundur, *Power System Stability and Control* (McGraw-Hill, 1994).

Separate from whether the variable of concern is voltage magnitude or angle, power engineers distinguish *steady-state*, *dynamic*, and *transient stability*. These categories overlap (sometimes confusingly) with the terms *small-signal* and *large-signal* stability in control theory.

Steady-state analysis concerns the physical limits for a feasible operating state. Specifically, what are the highest generator outputs and power flows that can be sustained? We will discuss these limitations with respect to power angle in Section 13.4.3, which involves generators, and with respect to voltage magnitude in Section 13.5, which focuses more exclusively on transmission lines.

Dynamic stability, most broadly, studies the response of a system to various types of *movement*, or changes to the system. These changes could be large or small disturbances, of brief or long duration, and they could be oscillatory (i.e., periodic) or single point-in-time events. In the power systems context, dynamic stability usually refers to the response to small but ongoing disturbances. This response is shaped by control loops on devices (e.g., generators or inverters) and their interactions, which are beyond the scope of this book.⁹ The central question is how the system collectively responds so as to diminish (i.e., *dampen*) rather than amplify these disturbances.

For example, dynamic stability studies oscillation *modes* of the grid, where power “sloshes” between geographic locations at some particular *subsynchronous* frequency (typically on the order of 1 Hz or less), akin to a resonance condition. These modes can be excited by *forced oscillations* due to some external input, such as malfunctioning generator controls or loads. Since the culprit may be hard to identify quickly, the relevant concern for operators is the amount of *damping* present in the system. Oscillation modes are an *emergent property* of the interconnected power system and must be studied empirically (see Section 16.2.4); there is no simple formula to predict them.¹⁰

Transient stability concerns the system’s ability to withstand and accommodate sudden large disturbances such as faults, loss of a transmission link, or failure of a large generating unit. The central question is whether the system can pass smoothly through this transient condition—either back to its predisturbed state, or to settle into a new and manageable steady operating state. We will address this topic in Section 13.4.4 for the context of generator angle.

13.4.2 The Concept of Stability

“Stability” is a rigorous application of the term from physics, where one distinguishes different types of equilibria, stable or unstable, that describe the tendency of a system to depart from or return to a certain resting condition in response to a disturbance. A simple mechanical analog is useful. Consider a smooth round bowl and a marble, as in Figure 13.2. Inside the bowl, the marble rests at the bottom in the middle. If we displace the marble from this resting point, by moving it up toward the rim of the bowl, its tendency is to return to the equilibrium location: it will roll around or back and forth in the bowl until it settles again at the bottom. This is a *stable equilibrium*. Gravity acts here as a *restoring force* that drives the system back toward its equilibrium state (i.e., pulls the marble to the bottom).

Now imagine the bowl turned upside down, with the marble precariously balanced on top. As long as the marble is situated precisely at the highest point, without so much as a breath disturbing it, it will stay. But the slightest displacement will make the marble roll off the side of the bowl. Gravity no longer acts as a restoring force, but to exacerbate the marble’s displacement away from equilibrium. This type of equilibrium is *unstable*.

⁹ Section 13.4.3 briefly refers to *damper windings* and Section 10.4.2 briefly mentions Power System Stabilizers (PSS) associated with automatic voltage regulators (AVR) on synchronous generators.

¹⁰ An excellent overview on grid oscillations is Jim Follum, Francis K. Tuffner, Luke A. Dosiek, and John W. Pierre, *Power System Oscillatory Behaviors: Sources, Characteristics & Analyses* (2017). PNNL-26375/NASPI-2017-TR-003. Available: naspi.org/node/629 (accessed February 2024).

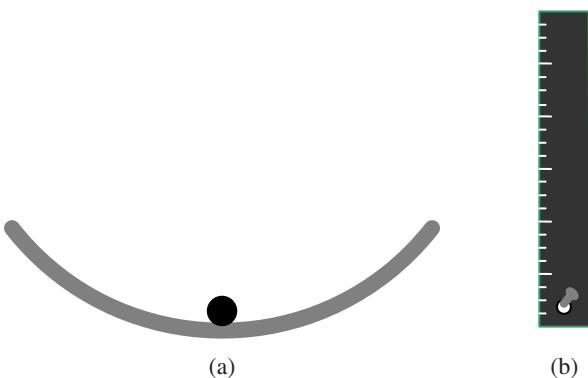


Figure 13.2 Stable and unstable equilibria: (a) bowl and marble in stable equilibrium; (b) ruler in unstable equilibrium.

Another simple example is a ruler with a hole at one end, supported by a nail on the wall. In its stable equilibrium, the ruler hangs down from the nail. Displace it by lifting the other end, and the ruler will swing back and forth, eventually coming to rest again in the vertical position. How do we make an unstable equilibrium with this system? Turn the ruler upside down so that it stands vertically above the nail. Again, if positioned carefully enough and left alone, the ruler will stay, but the slightest disturbance will cause it to swing back around.

Analogous to the marble or the ruler, what is “moving” in a power system is the instantaneous voltage at a given location. Stability analysis studies the forces that tend to restore or displace the voltage from an equilibrium condition.

13.4.3 Angle Stability

Angle stability focuses on the relative timing of the voltage maximum, expressed as the *voltage angle* or *power angle* δ . As we saw in Section 10.4.3 on controlling the real power output of synchronous generators, a generator’s power angle can move ahead or behind in time, and this movement is associated with an exchange of power between this and other generators, mediated by a net circulating current. In Chapter 12, we come to think of the power angle as a variable that, for any location in the system, indicates how much real power is being injected into or drawn out of the system at that point. In the context of stability analysis, we understand the power angle as a variable that couples different and remote parts of a power system to each other.

When considering steady-state stability, we are basically asking whether a particular operating configuration of an a.c. power system represents a stable equilibrium. In other words, is it possible to supply a given set of loads with a given set of power contributions from generators through a given network of transmission lines, and maintain synchronism among all components? *Synchronism* means that both the *frequency* and the *phase* of two or more oscillating components match. When synchronous generators are connected together, they must be spinning at precisely the same rate,¹¹ and they must also be in step with each other: that is, the zero crossings and peaks of their voltages should occur at the same time—or, more precisely, within a fixed, very small time interval δ of each other. Only in this way can all generators simultaneously contribute to feed power into a network.¹²

¹¹ This statement refers to the electrical, not the mechanical rotational frequency; the two are directly related by the number of magnetic poles (see Section 10.3.2).

¹² If the timing of voltage and current did not match up for each generator, it would have to alternately inject and absorb power over different portions of the cycle. In practice, this would cause overloading of parts of the generator windings, if the circuit breakers had not opened first to protect the machine from damage.

Synchronism also requires a stable equilibrium condition where there is a restoring force that tends to slow down a generator that has sped up, and to speed up a generator that has slowed down. Otherwise, synchronism could not be maintained, because the slightest disturbance would throw off individual generators and have them go at different speeds. Such a restoring force indeed exists; it was explained in terms of the power exchange between two generators in Section 10.4.3. The force results from the fact that a generator whose relative timing or *power angle* is ahead of others must supply additional power (thus tending to restrain the turbine more), whereas one whose power angle is behind supplies less power (thus relieving the restraint on the turbine). This interaction provides for a negative-feedback effect that serves to control and “hold steady” each generator.¹³

When we say that a power system is operating within a regime of steady-state stability, we mean that it is in a regime where these stabilizing, restoring forces exist. In the marble-in-a-bowl analogy, the marble being displaced represents the individual generator rotor, and the marble’s position is the generator power angle, δ , relative to a reference. What makes the generator system tricky to conceptualize is that this reference is itself rotating. Thus, “equilibrium” does not mean things stand completely still: rather, it describes the condition where the machine is spinning at a steady, synchronous frequency and a steady power angle, with a balance between mechanical power delivered from the turbine and electrical power injected into the grid.

Section 13.4.4 will examine how the generator might speed up and slow down relative to its synchronous speed, thus pulling the power angle ahead or behind its equilibrium value. This relative movement of δ is represented by the movement of the marble. Steady-state stability is primarily concerned with the size and shape of the bowl.

This shape depends on both on the fixed properties of the network and its operating state, including the differences in phase or power angles at various locations. This is because the effective exchange of power between generators depends on the relative timing of their voltages and currents. As we illustrate in more detail later, the strongest negative feedback or stabilizing interaction between generators occurs when their phases are very close together. As the difference between their phase angles grows, which corresponds to a greater difference in power generation and thus a greater transmission of power between them, the interaction, and thus the stabilizing effect, weakens. Thus, the question for steady-state stability analysis is, How much power can we transfer, and still maintain a sufficiently concave shape of the bowl?

To complicate things, we might be interested in a single generator and its behavior relative to the larger system (modeled as an *infinite bus*), or the transfer of power over some distance across a transmission line. In either case, the theoretical limit for power transfer from one point to another depends on the *impedance* in between. For a single generator connected to an infinite bus, the relevant quantity is its internal impedance, discussed in Section 10.7.3; over distance, it is the transmission line impedance. In either case, we will assume that there is negligible resistance and $Z = R + jX \approx jX$.

With that assumption, we may use the approximate mathematical expression derived in Section 13.5 for the transfer of real power P_{12} between two points on an a.c. system:

$$P_{12} \approx \frac{|V_1||V_2|}{X} \sin \delta_{12} \quad (13.1)$$

This power transfer is a function of the *power angle* δ_{12} , or difference between the voltage phase angle at each point, which also describes the relative angular position of a synchronous generator

¹³ This is called a “negative feedback” because the force acts opposite (or in the negative direction) to the displacement: when the generator speeds up, slow it down; when it slows down, speed it up.

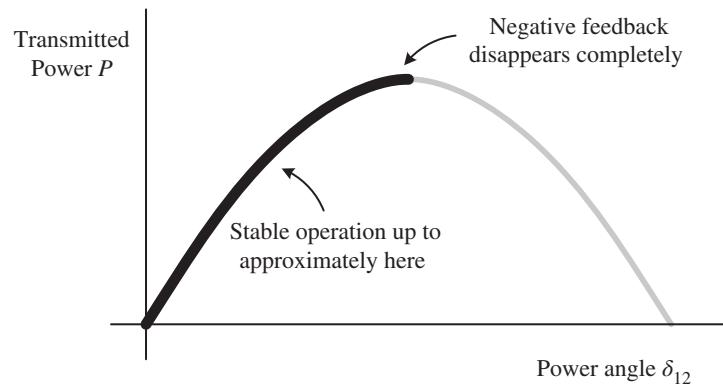


Figure 13.3 Power transmitted versus power angle.

there.¹⁴ $|V_1|$ and $|V_2|$ are the voltage magnitudes on either end, and X is the reactance of whatever lies in between points 1 and 2.

Figure 13.3 illustrates the dependence of power on δ_{12} in Eq. (13.1). If we take $|V_1|$, $|V_2|$, and X to be constant, the graph is simply a sine curve of the form $P_{12} = \text{const} \cdot \sin \delta_{12}$. Equation (13.1) thus establishes the theoretical limit for power transfer, which occurs when $\delta_{12} = 90^\circ$ and $\sin \delta_{12} = 1$.

However, this theoretical maximum power transfer is not a realistic steady-state operating condition. As soon as we consider the possibility of even small disturbances, a more restrictive stability limit on δ_{12} arises. The practical stability limit is based on the need for some negative feedback between generators at either end of a transmission line. This feedback diminishes as δ_{12} increases. We will explain this for the case of a single line connecting two machines, but the same reasoning extends to the entire network, where every transmission link may be examined for the δ_{12} between any two points where power is injected or withdrawn.

Let's review more carefully why the negative feedback between generators should depend on their difference in power angle. There are two approaches to illustrating this: mathematically or physically, referring back to the circulating currents responsible for exchanging power between generators as introduced in Section 10.4.3.

Equation (13.1) implies that we cannot maintain a steady operating state at exactly $\delta_{12} = 90^\circ$, because if, for any reason, δ_{12} were to fluctuate ever so slightly and increase beyond 90° , the power would actually *decrease*. This would represent an unstable equilibrium, since decreasing power would speed up the generator and lead to a further increase in δ_{12} . Therefore, it is unrealistic to operate either a generator or a transmission line at its theoretical maximum.

Furthermore, we note the slope of the sine curve, which is steep for small δ_{12} and gets flatter as δ_{12} increases. A flatter slope means that, for a given increment in δ_{12} , there is only a small increment in P_{12} . However, for a good, stabilizing feedback effect on the generator, we want the increment in P_{12} to be large.

This is analogous to the slope of the sides of the bowl with the marble: a weak restoring force would correspond to a shallow dish, which still tends to return the marble to the center, but not as quickly and reliably as a deep bowl (Figure 13.4). Therefore, from the stability standpoint, it is preferable to operate with a small δ_{12} where the slope of $\sin \delta_{12}$ and thus the incremental change in P is large.

This stability consideration weighs against the incentive to transmit larger amounts of power on a given line, which forces an increase of δ_{12} toward the shallower region. The transmission line

¹⁴ Here, we use δ for the power angle; in Chapter 12, this same angle is called θ .

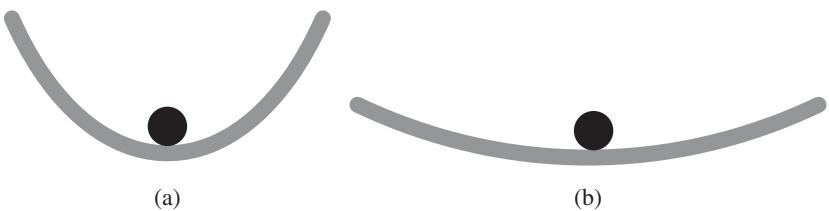


Figure 13.4 (a) Deep and (b) shallow equilibria.

could be upgraded by decreasing X , in which case one can achieve higher P for the same δ_{12} . The stability limit comes at the point where the slope of $\sin \delta_{12}$ is still steep enough for comfort, and leaves enough room for excursions due to possible disturbances, including large ones as considered in Section 13.4.4.

There is no clear, exact threshold for how much voltage phase angle separation is realistic or safe to sustain. Some rules of thumb quote 35° or 45° as a steady-state *angle stability limit* across a transmission line, but it depends on the system and the kinds of disturbances expected. For example, it might be found empirically that with large steady-state angle separations across certain transmission lines, small oscillations become less well damped. Consequently, assigning operational angle stability limits is a matter of judgment and risk analysis, not simply solving an equation.

Things get more complicated when considering a network with multiple power injections along the way. For example, a large synchronous grid can sustain angle differences greater than 90° over long geographic distances, because the points in between are stabilized by generators in proximity to each other. The stability limit as discussed above applies to each line connecting two nodes or substations. How far the angle can be safely stretched, say, from St. Louis to Boston, or from Denmark to Spain, is a more difficult assessment.

To return to a more physical explanation of negative feedback between generators as a function of power angle, we refer back to Section 10.4.3 that introduced the notion of a “difference voltage” due to a difference in the timing of the voltage maximum because one generator has pulled ahead (greater power angle) or fallen behind (smaller power angle). The difference voltage results in a change of the current in the armature windings of the respective generators, which can be considered a circulating current superimposed on the regular load current, as in Figure 10.20.

The key requirement is that this circulating current is timed to increase the load of the generator that is ahead of the other (Unit 1) and reduce the load of the one that is behind (Unit 2). In Section 10.4.3, we argued this was the case, on the implicit assumption that the difference in timing between the two voltages maxima was a small fraction of a cycle. The resulting circulating current between the two units is associated with an additional positive power output from the unit whose voltage phase angle is ahead, while reducing the load for the other.

But what if the difference in the timing between the voltage maxima—that is, the difference δ_{12} in the power angle—grows large? As δ_{12} increases, the magnitude of the difference voltage and thus the circulating current increases dramatically, suggesting that more power is transferred from Unit 1 to Unit 2. But the circulating current will also be shifted out of phase with the main voltage waveform. As a result, the instantaneous product of current and voltage is sometimes positive and sometimes negative, reducing the average power transferred. This is one way to think about the gradual loss of stabilizing effect as the difference in power angles increases.

The greater magnitude of the circulating current at large δ_{12} can also become problematic in terms of heating the windings. For all these reasons, the effective control of interconnected synchronous generators becomes more difficult with increasing difference in power angle, and it is not desirable to let the steady-state angle separation δ_{12} become too large.

13.4.4 Transient Angle Stability

In keeping with our marble-in-the-bowl analogy, while steady-state stability is concerned with the shape of the bowl, transient stability is concerned with how far we can displace the marble (i.e., the power angle) and be sure that it comes to rest again. Transient stability can be considered a subset of dynamic stability, but here we are concerned with major displacements of the power angle due to a temporary but potentially large disturbance. The key question is, how large a disturbance can a generator sustain and still return to an equilibrium? Depending on the type of disturbance, this might be the same equilibrium as before, or a new operating point based on changed external conditions. In either case, we also care about whether it will settle down in its equilibrium in a reasonable amount of time.

To explain dynamic stability, many engineering texts use a different mechanical analog—namely, a spring—which is more accurate mathematically in representing a *simple harmonic oscillator*. As the spring is stretched or compressed, it tends to return to its original shape, with some bouncing back and forth until it comes to rest in its equilibrium position. The linear displacement of the spring is analogous to the displacement of the power angle. For our purposes of intuitive understanding (and especially for readers who are not already familiar with the analysis of the spring system), let's stick with the marble in the bowl.

As the marble moves around in the bowl—say, if we were to release it near the rim and let it roll down, up again on the other side, and so on—we can describe its behavior in terms of energy: it possesses a certain amount of gravitational potential energy, depending on its height at any given moment, and it possesses a certain amount of kinetic energy, depending on its speed. The sum of kinetic and potential energy remains constant (owing to energy conservation), except that gradually, as the marble keeps rolling, this energy will be consumed by friction (turning it into heat). Friction thus acts as a *damping* force that slows the oscillation and allows the marble to finally come to rest at the bottom.

Similarly, the electric generator has a certain amount of energy, a restoring force, and a damping force. Mathematically, this situation is described by a power balance equation, which is a statement of energy conservation: at any instant, the power going into the spinning rotor must equal the power going out. Thus, in equilibrium, the mechanical torque from the turbine shaft equals the electrical power that is pushed out the armature windings through the magnetic field, plus a small amount of damping. Away from equilibrium, when there is excess mechanical power supplied from the turbine, the rotor speeds up; when the power supplied from the turbine is less than the electrical plus damping power, the rotor slows down.

This power balance can be written as the generator *swing equation*, which is a differential equation¹⁵ that implicitly describes the behavior of the power angle $\delta(t)$ as a function of time. This swing equation is conventionally written as

$$M\ddot{\delta} + D\dot{\delta} + P_G(\delta) = P_M^0 \quad (13.2)$$

The swing equation may be written in megawatts, or in per-unit quantities (see Section 8.7).

¹⁵ A differential equation is one that relates a variable (say, the position of an object) to its rate of change (say, the velocity and/or acceleration of the object). Such an equation can be written down by considering basic laws of physics and the various forces acting on the object. A “solution” to a differential equation is a function that states explicitly how the variable behaves (say, the object’s position as a function of time), where this function must satisfy all the criteria stipulated by the differential equation. Usually more than one function meets these criteria, but the solution can be further narrowed down by identifying *boundary values* that the function must take on at certain points in order to fit a specific situation. For the differential equation describing an oscillating object, the general solution is some sort of sinusoidal function.

Before we study the swing equation itself, let's briefly preview its solution. The power angle $\delta(t)$ that satisfies Eq. (13.2) will appear as a *damped harmonic oscillation*, described by a sinusoidal function and (we hope!) an exponential decay term. This oscillation will have some sub-synchronous frequency (say, on the order of 1 Hz), and should subside noticeably over the course of several seconds.

Figure 13.6 gives an example of the response of a generator to a sudden disturbance in the system, as measured by the voltage phase angle at its terminal relative to another location on the grid. Although the actual scenario depicted is more complicated than the ones in this chapter, the figure illustrates the characteristic shape of the damped harmonic oscillation as modeled and as measured empirically. Because frequency is the rate of change of angle, and because the oscillations are roughly sinusoidal (where cosine is the rate of change of sine), graphs of angle oscillation appear very similar regardless of whether the vertical axis shows angular position or frequency.

Now let's examine each term in the swing equation (13.2), from right to left. P_M^0 is the mechanical power input from the turbine, where the superscript 0 indicates the value at equilibrium (i.e., constant power generation at the equilibrium power angle δ_0). Regardless of any disturbance, we assume that P_M^0 does not change over the time period of interest—that is, the spinning turbine is pushing with a constant force because it simply can't be adjusted that quickly.¹⁶ This mechanical input power on the right must equal the total on the left-hand side because of energy conservation.

The electrical power output is represented by $P_G(\delta)$. The notation with parentheses emphasizes that the power generated P_G itself varies as a function of the power angle δ . This angle can be considered spatially as the rotor position, or temporally as the voltage phase angle. In either case, δ is defined within a *rotating reference frame*, as a *displacement* from a reference position, all spinning at the synchronous frequency.

The functional dependence $P_G(\delta)$ is none other than the relationship we previously encountered in Eq. (7.1) and Figure 7.15, and again in Section 13.4.3 and Eq. (13.1). In the context of modeling individual generator behavior relative to an infinite bus with an ideal fixed voltage, the impedance X in the denominator refers to the generator's own internal reactance (Section 10.7.1). For practical scenarios, the limiting factor is likely the impedance of the transmission line across which power from the generator is to be transferred. The relationship is shown again in Figure 13.5, where P_G , the real power generated, varies as the sine of δ , the voltage angle difference between the generator internal voltage and a chosen reference point on the other side of the impedance. Therefore, over a reasonable range of δ , as δ increases, so does P_G .

However—and this is crucial!— P_G will not continue to increase indefinitely for increasing δ . In fact, for extremely large δ , P_G eventually becomes negative. It makes sense that P_G should eventually become negative, because the power angle δ describes the position or timing of a generator *relative to others*, and because we are dealing with a cyclical motion. At some point, it will lead so far ahead of others that it appears to lag behind them. Thus, if the generator pulls ahead of others by too great a phase angle, at some point it will be producing less instead of more electrical power.

Physically, we can think of P_G as the restoring force that pushes back on the rotor through the magnetic field: the farther we displace δ , the harder the magnetic force pushes back—up to some point. At the equilibrium point δ_0 , where the forces are balanced and δ holds steady, P_G equals P_M^0 , which is seen graphically in Figure 13.5 as the intersection of the P_G curve with the horizontal line that marks the value P_M^0 .

¹⁶ Note that an inverter might behave differently. The analytic tools in this section all relate to traditional rotating machines.

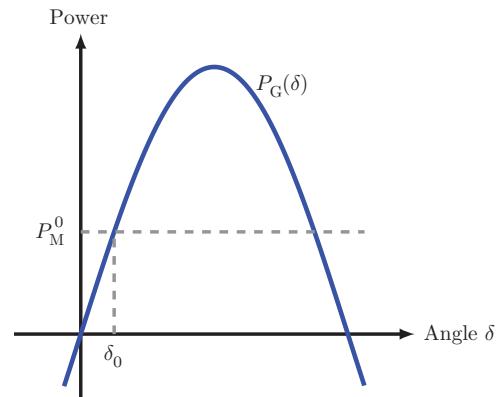


Figure 13.5 Power generated as a function of power angle.

The other two terms in the swing equation involve rates of change of the power angle. The single overdot indicates the rate of change (the first derivative) of δ with respect to time, which is analogous to velocity. Here, it corresponds to the rate of rotational displacement from the reference: when $\dot{\delta}$ is positive, the rotor's frequency is greater than the system frequency of 60 cycles (i.e., in the process of moving δ ahead of other generators); when $\dot{\delta}$ is negative, the rotor frequency is less than 60 cycles and in the process of falling behind.

The constant D is a measure of the damping force, whose tendency is to resist any changes in δ . In many physical systems, the term $D\dot{\delta}$ simply represents power absorbed by friction. In synchronous generators, damping is deliberately introduced with *damper windings*. Owing to clever geometry, these windings produce a torque that acts to oppose changes in the angular rotor position relative to the synchronous rotation; the details are beyond our scope. Like other introductory texts, we will neglect the damping term for purposes of mathematical analysis, and then sheepishly assert that some amount of damping must be included in a realistic generator. Fortunately, dynamic stability analysis that ignores $D\dot{\delta}$ will err on the conservative side, as any amount of damping will tend to reduce generator oscillations.

The double overdot indicates the rate of change of the rate of change of δ (the second derivative), which represents the *acceleration*, or change of speed, of the rotor. The constant M is a measure of the generator's *inertia*, whose effect is to resist changes in rotational speed. The power that goes into speeding up or that comes out of slowing down the generator rotor is represented by $M\ddot{\delta}$. This term is crucial for the overall behavior of a generator and consequent system stability.

Since every term in the physical swing equation has dimensions of power (energy per unit time), and $\ddot{\delta}$ has inverse seconds squared, M should have dimensions of energy \times time (also known as *angular momentum*).¹⁷ In physical units, it is given by

$$M = J\omega_s \quad (13.3)$$

where J is the machine's rotational inertia that depends on its size and shape (specifically, the distribution of rotating mass relative to the axis of rotation), and ω_s is the synchronous frequency in radians per second.

In case the machine has more than two poles (Section 10.3.2), it is necessary to define the *mechanical* rotational frequency $\omega_{m,s}$ and use it in lieu of ω_s when the physical movement matters, as in Eq. (13.3). Likewise, we must use the *mechanical* angular acceleration $\ddot{\delta}_m$ when considering the gain or loss of kinetic energy, while the electrical power injected $P_e(\delta)$ remains a function of the *electrical* angular position. An example at the end of this section will illustrate this subtlety.

¹⁷ The units of radians are blithely left to disappear, since they don't have physical dimension.

The swing equation is most often written in per-unit quantities, for which we simply divide the entire equation by the generator's rated power in MVA, S_{Base} (see Section 8.7). This format not only eases notation, but allows for a more consistent and intuitive comparison among generators of different sizes. Here, M is defined in terms of the *per-unit inertia constant* H :

$$M = \frac{H}{\pi f_s} = \frac{2H}{\omega_s}$$

where H represents the rotational kinetic energy of the machine at synchronous speed, divided by its rated power:

$$H = \frac{1}{2} \frac{J \omega_s^2}{S_{\text{Base}}}$$

The per-unit H thus has units of seconds, or megajoules per megavolt-ampere. As noted in Section 11.1.1, the intuitive interpretation is that H tells us how long a generator could keep pushing its rated load based on inertia alone, if the prime mover suddenly disappeared. That interpretation is technically unrealistic, but conceptually in the right spirit. Typical per-unit inertia constants are in the single digits of MJ/MVA and vary over a fairly narrow range, for machines of rather different sizes.

We now return to the swing equation as a whole. First, note how it accounts for the equilibrium condition: When δ just sits at the equilibrium value δ_0 that makes P_G equal to P_M^0 , there is no change of δ with respect to time, meaning that the damping and acceleration terms in the equation are zero. The generator keeps spinning and supplying electrical power at a constant rate, which in our rotating reference frame means that the whole situation is at rest.

But when analyzing dynamic stability, we want to know how the generator will respond to a disturbance of some sort, and whether it will be able to settle again in its equilibrium. In other words, what will happen if δ is somehow displaced from equilibrium? The essence of this analysis is to determine just how far δ can be displaced before there is trouble.

Classic cases for study involve the interruption of transmission links. For example, if one of several transmission lines in a path suddenly goes out of service, the impedance of the remaining path increases. This new impedance value changes the shape of the $P(\delta)$ curve (making it smaller). Consequently, the operating point for the same mechanical power P_M^0 will correspond to a different (greater) value of δ on the new curve. The question is, will the generator settle at this new value?

A simpler case that we'll analyze is where the transmission line is momentarily interrupted, but quickly reconnected. This type of event occurs, for instance, when a reclosing circuit breaker operates. Suppose the transmission link is a generator's only connection to the grid. Thus, during the time period where the link is interrupted, the generator cannot send out electric power. But because the time interval in question is very short—perhaps half a second—the steam turbine output cannot be adjusted. The turbine therefore continues to push the rotor with constant mechanical power input, P_M^0 . Because no power goes out in the form of electricity, all of P_M^0 goes into accelerating the rotor (except for a tiny amount to overcome friction).

During this interval where the generator is disconnected from the grid, the rotor gains momentum, which means that the power angle δ increases, as does its rate of increase $\dot{\delta}$ (because it continues to *accelerate* under the turbine's torque for the duration of the disconnect). The generator thus acquires a certain amount of excess energy, which is just the accumulation of turbine power over that time interval. Aside from any dissipation by damping, this excess energy manifests as kinetic energy of the rotor, which is now spinning at a higher than normal frequency.

The amount of this accumulated excess energy is crucial for determining stability. Up to a certain amount, it can be dissipated; more than that, the generator cannot return to equilibrium. Transient

stability analysis is concerned with determining that critical amount of excess energy. Naturally, the longer the generator is accelerating in its disconnected state, the more kinetic energy it will build up. Therefore, the problem is often stated in terms of the length of the time interval of disconnection. In other words, how long can the generator be disconnected before it gains so much speed that it will not return to normal when reconnected?

This question can be answered by turning to the solution of the swing equation. This equation was derived from the general principle of energy conservation, but it also dictates very specifically how δ may evolve as a function of time. Its mathematical solution says that, if δ is displaced a bit and then let go, it will oscillate back and forth, going alternately ahead and behind of δ_0 . Owing to the damping force, the oscillation should diminish over time and δ should eventually settle down into its equilibrium, δ_0 .

To see how this oscillation comes about, consider what happens at the end of the transient disturbance, when the generator is reconnected. As the transmission link is reestablished, the speeding generator can relieve itself of its excess energy into the grid. Indeed, it suddenly encounters a $P(\delta)$ that is very large, corresponding to the now very large δ , and that exceeds the turbine input power P_M^0 . The rotor decelerates. However, even after the rotor has begun to decelerate, there is a period during which δ still increases because the rotor is still spinning faster than the other generators in the system. This is analogous to the marble that has been given a good push and rolls up the side of the bowl, even though it is already under the influence of gravity pulling it back. The rotor continues to decelerate until δ is less than δ_0 , at which point $P(\delta)$ is less than P_M^0 and the rotor begins to accelerate again. As before, though, δ continues to decrease despite the fact that it already has positive acceleration, which is why it overshoots δ_0 until it reaches a minimum where it turns around again. This movement would continue back and forth indefinitely, save for the damping force that slows the motion of δ and causes the excursions to gradually diminish until δ settles at δ_0 .

If the initial displacement of δ was too far, however, it will not return. Essentially, this is the point where the additional power that the generator supplies to the grid when δ is ahead is not sufficient to slow the generator back down, because it has already built up too much momentum, or acquired too much energy. Such a point exists because $P(\delta)$ begins to *decrease* once δ gets too large, and indeed eventually becomes negative. By analogy, the marble has been given too big a push, so that the force of gravity can no longer confine it below the rim of the bowl.

This situation can be more specifically analyzed in terms of the exchange of potential and kinetic energy as the object (marble or rotor) oscillates. For the marble in the bowl, we can easily see that a certain amount of energy corresponds to the marble being confined to the bowl, while a greater amount of energy would imply that the marble jumps over the rim. Notably, it does not matter whether we are talking about kinetic or potential energy, since either can propel the marble out of the bowl: if the marble has too much potential energy, this means it will be located too high, past the rim of the bowl; if it has too much kinetic energy, it will move too fast and overshoot the rim on its next upward roll. We can thus state that the marble will stay inside the bowl as long it has no more than a certain maximum total energy, which is the sum of potential and kinetic.

For the generator whose power angle is oscillating or swinging, there is an analogous limit of total energy, which can also be described as the sum of a potential and kinetic energy. However, while this nomenclature establishes an easy mathematical analogy with other physical systems, it can also be confusing. Therefore, while keeping in mind the rest of the grid, let us first consider what physically happens to the energy during the generator's oscillation.

Because the turbine power inputs and loads throughout the network do not change, and because energy is conserved, the only place where energy can increase or decrease is in the rotational kinetic energy of the generator rotors. During the interval that the generator in question is disconnected

and speeding up, other generators—for simplicity, let's say just one other generator—elsewhere in the system is supplying the extra load and is therefore slowing down. After the connection is reestablished, the two generators now have a symmetric power imbalance: one is too fast, the other too slow. The ensuing oscillation is the exchange of energy back and forth between these two generators, which alternately speed up and slow down until they again share the load according to their set points. Figure 13.6 illustrates such an oscillation.

The individual generator that is immediately affected by the transient disturbance alternately gains and loses rotational kinetic energy, as its rotational speed increases and decreases. It has a maximum amount of kinetic energy when the rotor is spinning fastest, meaning that δ is increasing most rapidly. Perhaps counterintuitively, this point does not coincide with the point at which δ itself is maximum (where δ actually holds steady momentarily and the rotor spins at its nominal 60 cycles, despite being displaced), but rather when the rotational frequency or $\dot{\delta}$ is maximum, while δ increases and passes δ_0 . This point corresponds to the bottom of the bowl where the marble has maximum speed.

On the way back, as δ decreases and passes δ_0 , the rotor is physically moving the slowest. If we were describing the rotational kinetic energy of the generator in strict physical terms, we would say that it is at a minimum at this instant. However, for the purpose of analyzing the oscillation, we construe a mathematical quantity called “kinetic energy” that does not care whether the speed *relative* to the nominal 60 cycles is positive or negative, as if the power angle itself (which is really only a label marking the position of the rotor relative to that of other generators) were an actual physical object in motion. Given by $\frac{1}{2}M\dot{\delta}^2$, this quantity is mathematically analogous to the kinetic energy $\frac{1}{2}mv^2$ of a rolling marble (where m is mass and v is velocity), and it is the same regardless of the direction in which the marble is rolling. Therefore, the so-called kinetic energy is again at a maximum as δ passes δ_0 in the opposite direction, when $\dot{\delta}$ is greatest in the negative direction.

Similarly, we can define a “potential energy” for the rotor, which is analogous to the gravitational potential energy of the marble. When δ is equal to δ_0 , this “potential energy” is zero (the marble is at the bottom of the bowl). As δ is displaced, in either direction, the generator acquires “potential energy.” We can think of this potential energy as the accumulation of restoring power, or work the

Unit Trip 2011/05/30 03:03:00 GMT. UT Pan Am Relative to U.T. Austin.

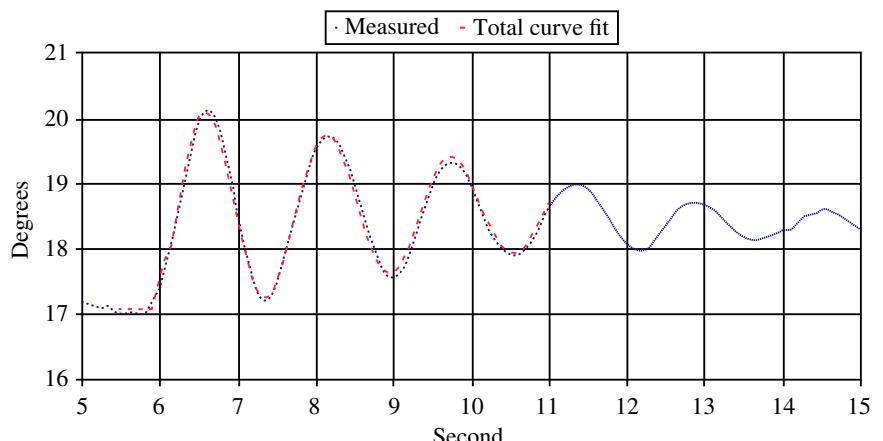


Figure 13.6 Damped harmonic oscillation in $\delta(t)$, as modeled and measured between two locations in Texas, illustrating generator swing and ringdown after the sudden loss of another generator. Source: Courtesy of Mack Grady.

generator has performed by squeezing excess power into the grid. As δ is displaced farther, we can imagine more and more capacity built up to push δ back in the opposite direction, with a maximum potential energy at the point of maximum δ . The restoring power thus accumulated is the difference between the electrical power output and the mechanical power input. This difference is represented in Figure 13.7 as the portion of the $P_G(\delta)$ curve that extends above the line representing P_M^0 . Mathematically, the potential energy is the *integral* of the restoring power over δ , or the area under the curve between δ_0 and the given δ . Note that the units on the horizontal axis are degrees, which represent time, so that the area (power·time) has dimensions of energy. The maximum potential energy occurs at the point of farthest displacement of δ , just like the marble has maximum potential energy at its highest point. The curve of cumulative potential energy as a function of δ , conventionally labeled $W(\delta)$, is shown in Figure 13.7. This curve is analogous to the bowl.

It may be counterintuitive that the generator ought to have a maximum of potential energy at both maximum and minimum δ (just as it has maximum kinetic energy at both maximum and minimum physical speed). At maximum δ , it has the greatest capability of doing work on other generators, that is, to relieve their load by carrying extra power and thereby slowing down, sacrificing its own “lead” in δ . At minimum δ , the situation is reversed, where our generator now has the greatest capability to absorb the extra work of others. For the purposes of stability analysis, these two situations are symmetric and are given the same label.

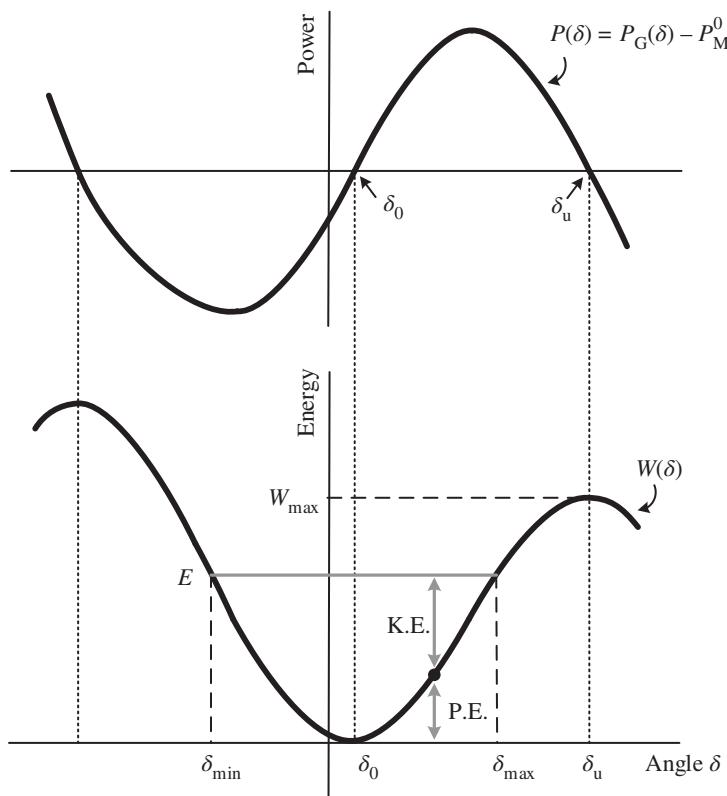


Figure 13.7 Restoring power and “potential energy” $W(\delta)$, where K.E. = kinetic energy; P.E. = potential energy.

The peak of the potential energy curve in Figure 13.7—specifically, the right peak, if we are concerned about a forward displacement of δ —indicates the maximum amount of restoring work that can be done on the generator to bring it “back in line.” It is labeled W_{\max} . Physically, this means the maximum energy that the generator can dump into the grid by running ahead of others. (Conversely, the peak on the left shows the maximum amount of energy that the generator can absorb from others if it is behind.) This corresponds to the accumulated restoring power from δ_0 to the upper limit, labeled δ_u , or the area between the $P_G(\delta)$ curve and P_M^0 .

Finally, these definitions provide us with a concise way to state the transient stability criterion. During the transient condition, the generator acquires both potential energy (because δ is displaced) and kinetic energy (because δ is in the process of changing). The generator is transient stable if the total energy acquired during the transient period is no more than the maximum amount of energy that can be gotten rid of, W_{\max} . The beauty of this articulation is that it is very general and assumes nothing about the particular transient disturbance.

There are two ways of illustrating this condition graphically. The first is on the $W(\delta)$ curve, where the excess energy acquired by the generator is represented in terms of height. It gains height as it moves out along the curve with increasing δ . During the oscillation, the total energy remains constant (save for damping), which is indicated by the horizontal line E . This total energy is composed of a potential energy component (the height of the W curve at a given δ , labeled P.E.) and a kinetic energy component (the difference between total and potential energy at the same δ , or the distance between the W curve and the horizontal line E , labeled K.E.). Stability requires that the horizontal line E must not go above W_{\max} . Note that this does not allow us to displace δ as far as δ_u by the end of the transient period because the displacement will also entail kinetic energy (i.e., still be in the process of increasing) at the moment that it is “let go.” As a result, given a certain total energy E , δ will increase up to δ_{\max} and then swing back to δ_{\min} before hopefully settling at δ_0 . The stability limit is where δ_{\max} will never exceed δ_u , which is to say that the combination of potential and kinetic energy always remains less than W_{\max} .

This representation establishes a perfect analogy to the condition that the marble must not overshoot the rim of the bowl. In particular, we must not let go of the marble near the rim if it has too much upward velocity! It also illustrates the fact that instability can result from excessive slowing down of the generator, that is, moving it out on the left side of the W curve, which would cause it to then oscillate and overshoot the peak on the right. This situation corresponds to a sudden, enormous load holding back the generator, which could be brought about by the failure of another large generator in a small system.

The other representation refers to the graph of electrical power $P_G(\delta)$ and turbine power P_M^0 , shown in Figure 13.8. This representation provides us with a specific δ as the stability limit. Here, the stability criterion is articulated as the *equal-area criterion*. It states that the generator is transient stable if the area below P_M^0 , between δ_0 and the displaced δ_T at the end of the transient, is no greater than the area between $P_G(\delta)$ and P_M^0 up to their intersection point, which occurs at δ_u and represents the maximum δ for which there is still any restoring force at all. The first area represents the amount of excess energy acquired by the generator during the transient period (where the full turbine power is being absorbed); this is why it is called the “acceleration area.” The second area, called the “deceleration area,” represents the cumulative deceleration power, or the total amount of energy that the generator can dump into the grid, minus the amount to which it is already committed due to its displacement of δ . In other words, there must be enough deceleration power left to absorb the excess kinetic energy once the potential energy has been accounted for.

For a given generator, the manufacturer will specify a curve $P_G(\delta)$ as a basis for this type of analysis. While each machine behaves according to its own swing equation, the aggregate

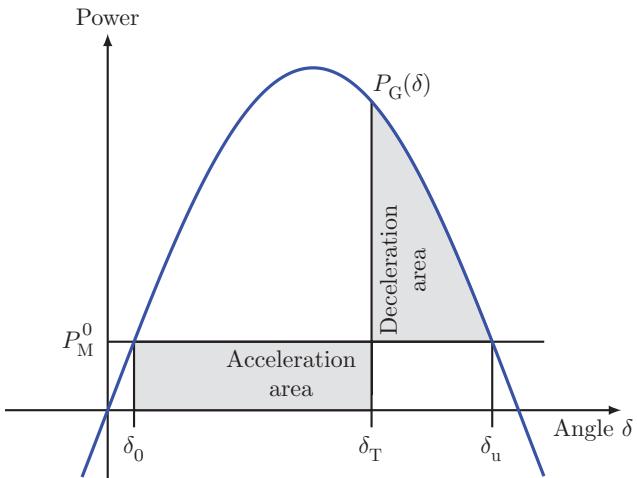


Figure 13.8 Equal area criterion.

phenomenon of a.c. grid stability involves interactions between and among all devices across the entire network (defined by transmission line impedances), with resulting oscillations in voltage angle and frequency observable at every location in the grid. Empirical measurements of oscillations caused by multiple generators in an a.c. network swinging against each other in response to sudden disturbances are shown in Figures 13.6 and 16.5.

This type of realistic dynamic behavior at the system level is far more complicated than anything that can be calculated or predicted with tools like the ones in this chapter, and remains a challenge even in state-of-the-art power systems analysis. Direct measurements of local voltage phase angles with PMUs (Section 16.2.4) have been essential for producing and validating useful models, to better evaluate and predict the complex system dynamics of large networks with significant long-distance power transfers.

An additional analytic challenge is introduced by the replacement of synchronous generators with inverter-based resources (Section 14.4), which do not have intrinsic rotational inertia like the constant M in the swing equation. With suitable control algorithms and some energy storage, power electronic devices can emulate synchronous generators to produce *synthetic* inertia, in the interest of keeping the system behavior predictable within traditional operating rules. Inverters are also capable of producing dynamic responses that serve *better* than legacy rotating machines to stabilize the grid, for example, with more active damping of oscillations. Control strategies for system stability promise to remain an interesting area for research and development in the years ahead.

Example

A 50-Hz, four-pole generator is rated 200 MVA and has a per-unit inertia constant $H = 5.0$. Find the electrical and mechanical angular acceleration if the machine is disconnected from the grid while at synchronous speed and full rated output.

The electrical and mechanical synchronous angular frequencies are

$$\omega_s = 50 \text{ Hz} \cdot 2\pi \text{ rad} = 314 \text{ rad/s}$$

$$\omega_{m,s} = \frac{1}{2} \omega_s = 157 \text{ rad/s}$$

The rotational inertia J is

$$J = \frac{2 H S_{\text{rated}}}{\omega_{m,s}^2} = \frac{2 \cdot 5 \text{ MJ/MVA} \cdot 200 \text{ MVA}}{(157 \text{ rad/s})^2} = 0.081 \text{ MJ} \cdot \text{s}^2 = 8.1 \times 10^4 \text{ kg} \cdot \text{m}^2$$

The swing equation without damping can be written as

$$P_m - P_e(\delta) = M \ddot{\delta}_m = J \ddot{\delta}_m \omega_{m,s}$$

Thus we can solve for

$$\ddot{\delta}_m = \frac{P_m - P_e}{J \omega_{m,s}} = \frac{200 \text{ MVA}}{0.081 \text{ MJ} \cdot \text{s}^2 \cdot 157 \text{ rad/s}} = 15.7 \text{ s}^{-1} \text{s}^{-2} \text{s} = 15.7 \text{ rad/s}^2$$

Since every radian of mechanical rotation corresponds to two radians of electrical phase advancement, the electrical angular acceleration is

$$\ddot{\delta} = 2 \ddot{\delta}_m = 31.4 \text{ rad/s}^2$$

With this information, we can determine the evolution of the power angle $\delta(t)$ as it becomes increasingly displaced during the transient imbalanced condition. For the given scenario this is easy to do, since the acceleration remains constant throughout. We have

$$\dot{\delta}(\tau) = \int_0^\tau \ddot{\delta} dt = \ddot{\delta} \tau$$

and

$$\delta(\tau) = \int \int_0^\tau \ddot{\delta} dt = \int_0^\tau \dot{\delta}(t) dt = \int_0^\tau \ddot{\delta} t dt = \frac{1}{2} \ddot{\delta} \tau^2$$

For example, after 10 cycles (one-fifth of a second at 50 Hz), the rotor will have advanced from δ_0 by

$$\delta(0.2) = \frac{1}{2} 31.4 \text{ rad/s}^2 (0.2 \text{ s})^2 = 0.63 \text{ rad} = 36^\circ$$

Does this spell trouble? How long before it's too late to reconnect the machine? To answer these questions, we need to know the $P_e(\delta)$ function, which informs us about the initial angular position δ_0 and about the kinetic energy absorbed by the grid after our machine reconnects and decelerates. Let's suppose this function is given by

$$P_e(\delta) = 5.0 \sin \delta \text{ p.u.}$$

Referring to Figure 13.8, we can apply the equal area criterion as an estimate for how late is too late. This will require us to perform an integral to account for the changing value of $P_e(\delta)$.

Calling δ_T the angle at the moment when the machine reconnects, our general expression for the acceleration area AA is

$$AA = P_M \cdot (\delta_T - \delta_0) = 1.0 \text{ p.u.} (\delta_T - \delta_0)$$

where the original position δ_0 in equilibrium at full rated power is given by

$$1.0 \text{ p.u.} = 5.0 \sin \delta_0$$

$$\delta_0 = \sin^{-1}(0.20) = 0.20 \text{ rad} = 11.5^\circ$$

The deceleration area DA is determined by the power absorbed by the grid (in excess of P_M) between δ_T and the point of no return, δ_u :

$$\begin{aligned} DA &= \int_{\delta_T}^{\delta_u} (P_e(\delta) - P_M) d\delta = \int_{\delta_T}^{\delta_u} (5.0 \sin \delta - 1.0) d\delta \\ &= \int_{\delta_T}^{\delta_u} 5.0 \sin \delta d\delta - 1.0(\delta_u - \delta_0) = -5.0(\cos \delta_u - \cos \delta_T) - 1.0(\delta_u - \delta_T) \end{aligned}$$

From inspection of Figure 13.8 and symmetry, we find

$$\delta_u = \pi - \delta_0 = 2.94 \text{ rad}$$

Now we can solve for the maximum allowable δ_T by setting the two areas equal, $AA = DA$:

$$1.0(\delta_T - \delta_0) = -5.0 \cos \delta_u + 5.0 \cos \delta_T - \delta_u + \delta_T$$

Some rearranging gives

$$\cos \delta_T = \cos \delta_u + \frac{1}{5}(\delta_u - \delta_0)$$

With the above values for δ_0 and δ_u we have

$$\cos \delta_T = -0.52 \quad \text{and} \quad \delta_T = 2.02 \text{ rad} = 116^\circ$$

Note that it is okay for δ_T to *momentarily* overshoot the stability limit of 90° , given the substantial deceleration area still ahead under the $P_e(\delta)$ curve, and the fact that P_M was fairly small in proportion.

To find the amount of time it takes to arrive at δ_T , we return to

$$\delta_T = \frac{1}{2}\ddot{\delta} T^2$$

and solve for

$$T = \left(\frac{2\delta_T}{\ddot{\delta}} \right)^{1/2} = \left(\frac{2 \cdot 2.02}{31.4 \text{ rad/s}^2} \right)^{1/2} = 0.36 \text{ s} \approx 18 \text{ cycles}$$

So, in the above example, the generator should still be able to recover after 10 cycles—but the recloser had better act soon!

13.4.5 Voltage Stability

Up to here, we have taken “stability” to refer to stability of the voltage phase angle or power angle (*angle stability*). The stability of the voltage magnitude, referred to as *voltage stability*, is a less common day-to-day operational concern, but an important issue in its own right. Voltage magnitude is related to both real and reactive power, but the dependencies are analyzed separately and inform different sets of actions. Consistent with the “decoupling” approximation in power flow analysis (Section 12.4.6), reactive power is typically the more sensitive quantity with regard to voltage magnitude. In other words, an increase in reactive power Q injection or demand at a specific location tends to have a stronger effect on the local voltage magnitude than real power P , assuming transmission lines are dominated by inductive reactance.¹⁸ Unless stated otherwise, “voltage support” almost always refers to reactive power Q .

Recall that angle stability hinges on the condition that power output increases with power angle in a generator. Likewise, in a stable system, if we increase power injection, it should tend to raise the voltage magnitude at that location or bus in the network. The condition of *voltage instability* implies that no amount of real or reactive power injection is able to raise the voltage.

Like the power angle, voltage magnitude can exhibit an oscillatory behavior when displaced from equilibrium. Usually, departures from equilibrium are small and expected to dampen out very quickly, though sometimes continuing oscillations are observed after major disturbances. Voltage oscillations can propagate far, and larger geographic size may make synchronous power systems more vulnerable to voltage oscillations. Because so many different interacting components are involved, these effects on a large scale are very difficult to model.

¹⁸ Distribution lines with a small X/R ratio behave somewhat differently; see Section 12.6.

The explicit mathematical relationship between power transfer and voltage magnitude is treated in Section 13.5. It is more complicated than the relationship with voltage angle $P(\delta)$, and offers multiple interpretations of “voltage stability limits” depending on context. We revisit voltage stability in the context of P - V and V - Q curves in Sections 13.5.1 and 13.5.2.

Luckily, true voltage instability where voltage and power can no longer be controlled is rare in practice. It is generally preceded by a condition of low voltage. Even before a theoretical point of instability is reached, attempting to operate a system at a low voltage may cause other problems. For example, generator field excitations may increase, causing generator fields to be overloaded and then deliberately reduced or the generator tripped; transmission lines may trip; or transformer tap changers and voltage regulators may attempt to restore load voltage to normal and thereby further increase the load power. Angle instability may also result. Finally, a smaller or larger part of the system may be lost completely. This is the condition of *voltage collapse*.

If operators find empirically that they are having trouble maintaining voltage levels, it can be difficult to ascertain just how far away they actually are from voltage collapse—like walking near a cliff in the fog. This metaphorical cliff is known as the *nose* of the power-voltage curve, as depicted in Figure 13.10. Offline engineering analysis aims to establish the parameters for these power-voltage curves under different sets of operating conditions, so that operators know how much *margin* in megawatts there is between a given operating point and the nose. In routine practice, operators try to maintain the bus voltage magnitudes within relatively narrow ranges, with plenty of margin. A standard range is $\pm 5\%$ of nominal voltage (i.e., between 0.95 and 1.05 p.u.), and sometimes up to $\pm 10\%$.

As discussed in Section 10.4.4, the primary tool for managing the voltage magnitude profile across a network is to recruit MVARs from generators at various locations. In addition, many buses or substations have other dedicated resources at their disposal for the explicit purposes of maintaining the desired voltage magnitude. These include capacitor banks and static VAR compensators (see Section 7.4).

Generally speaking, voltage stability is enhanced, and the associated limit on real power that can be sent to the load is increased, by injecting reactive power near the load. Therefore, a power system can be especially vulnerable when operating at a low voltage if it is also low on *reactive reserves*, or reactive power generation that can be recruited to raise voltage magnitude. Such a situation preceded, for example, the August 14, 2003 blackout in the Northeastern U.S.¹⁹ The postmortem analysis showed that there was no voltage collapse, but that the voltage magnitude at various points in the system had continued to drop as multiple transmission line outages occurred, and the system was no longer in a secure state. The low-voltage condition led up to the blackout, which was finally precipitated by a cascade of protective relays tripped by a low voltage relative to the current.²⁰

Thus, analogous to the term *security* as it refers to the ability to generate and deliver real power (Section 13.3), there is a concept of *voltage security*, or width of the operating envelope with respect to voltage control. A system in a secure state can tolerate one or more worst-case disturbances (such as the loss of a generator or transmission line) without triggering a chain of events that will lead to a loss of load.

¹⁹ The very readable and educational Final Report by the U.S.–Canada Power System Outage Task Force can be found, among other places, at <https://www.energy.gov/sites/default/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf> (accessed February 2024).

²⁰ These types of relays are known as *distance relays* (because the low voltage/high current condition is characteristic of a fault in the distance) or *impedance relays* (because the ratio of voltage to current has units of ohms); see Section 7.5.

13.5 Power Transfer Limits

A formal or quantitative treatment of voltage stability rests on the analysis of power transfer across transmission lines. As introduced in Section 7.3, there are physical limits to the amount of power that can be transferred across transmission lines. In practice, the binding constraint is often the thermal limit based on I^2R heating. Less obvious limitations based on the relationships between sending and receiving end voltages come into play especially for longer lines.

Intuitively, we can appreciate that there must be a physical power transfer limit by considering that a transmission line's impedance itself acts as a load, even if it is purely inductive. Imagine connecting two resources (say, a generator and a load) together at the same physical location, with no line impedance in between. There is no limitation on power transfer whatsoever (other than the resources' internal constraints, which are not at issue here). But when we separate the two buses and introduce some finite impedance between them, that impedance becomes a load in series with the intended load. The longer the line and the greater its impedance, the more it acts to constrain the total current, and the greater the proportion of power delivered to the line versus the intended load.²¹

Formally, the limitations on power transfer are derived from the relationship between complex power and complex voltage. When considering these relationships, it is not always obvious which is the independent variable, or how it is being controlled. In a real transmission network, some combination of voltage and power control strategies will interact across multiple nodes. For purposes of a tractable analysis, we may choose one of several idealized representations.

The first is a *radial system*, where one end of the transmission line is simply a passive load that does not act to control voltage. This framing harbors another choice: what is the dependence of load on voltage? For example, do we assume the load has a fixed impedance, or that it consumes a specified amount of real and reactive power? The reality is likely somewhere in between (Section 6.3.1). If we assume a fixed P and Q for the load, the receiving end voltage depends on the power demand, the sending end voltage, and the line impedance.

In the radial framework, with no control over voltage at the load bus, we are stuck with the voltage drop caused by the line impedance. In the attempt to increase power delivery to the load, the load would have to draw an increasing current. But the greater the current, the greater the voltage drop across the line, and thus the lower the voltage at the receiving end. Because of the nonlinear relationship (power depends on voltage squared), there comes a point not only of diminishing returns, but *reversal*, where the load receives less power the harder we try. This possibility of reversal gives rise to the voltage stability problem of Section 13.4.5 and is illustrated in the power flow example in Section 12.4.5.

Alternatively, we may assume that some active sources or voltage regulating devices fix the voltage at both ends (as illustrated, e.g., in Figure 3.21). This scenario is operationally more sophisticated, but computationally much easier (see Section 12.2.3). In this case, the power flow is determined by the sending and receiving end voltages and the line impedance. Power transfer will be limited by the simple fact that voltages at either end of a transmission line cannot be infinitely large, or separated in phase angle by more than 90° .

The assumption that voltage magnitude is controlled by active resources throughout a network (say, within $\pm 5\%$) is what often justifies taking the bus voltage magnitudes $|V_1|$ and $|V_2|$ as fixed parameters while focusing on the relationship between voltage phase angle difference δ_{12} and real power transfer, as in Section 7.3.2. It is important to understand, though, that when something

²¹ This is analogous to the rheostat example from Section 6.1.1

is acting to stabilize the bus voltage near the load, this resource—whatever its nature—will be physically required to deliver some power locally in order to perform its job. It is not possible to “source” the power to the load from across the transmission line.

To estimate the power transferred from sending to receiving end of a transmission line as a function of voltages, the short line approximation (Section 9.3.3) is generally adequate. We can use it to derive theoretical limits on power transfer that need not be exact in order to provide insight and some practical guidance.

Let's consider a short transmission line with series impedance Z . We choose subscripts 1 and 2 for voltage and current at the sending and receiving end, respectively, and call S_{12} the complex power injected into the transmission line from Bus 1. Keeping a consistent notation, we call S_{21} the power injected at Bus 2. This makes the power *received* at Bus 2 from the transmission line $-S_{21}$. There is nothing asymmetrical about the situation; power could in fact be transferred in either direction. We will write down power S_{12} injected at Bus 1 and $-S_{21}$ received at Bus 2 (not for any physical reason, but just because it is more intuitive to imagine a sending and a receiving end).

Complex power for each bus is written as in Chapter 3, where current is expressed in terms of voltage drop and line impedance:

$$\begin{aligned} S_{12} &= V_1 I_1^* = V_1 \left(\frac{V_1 - V_2}{Z} \right)^* \\ -S_{21} &= -V_2 I_2^* = -V_2 \left(\frac{V_2 - V_1}{Z} \right)^* \end{aligned} \quad (13.4)$$

Importantly, the current at either bus must be the same: $I_1 = I_2$. Using $V_1 = |V_1| \angle \delta_1$, $V_2 = |V_2| \angle \delta_2$ and $Z = |Z| \angle \theta_Z$, collecting terms, and noting the sign change of angles due to the complex conjugate, we obtain:

$$\begin{aligned} S_{12} &= \frac{|V_1|^2}{|Z|} \angle \theta_Z - \frac{|V_1||V_2|}{|Z|} \angle (\delta_1 - \delta_2 + \theta_Z) \\ -S_{21} &= -\frac{|V_2|^2}{|Z|} \angle \theta_Z + \frac{|V_1||V_2|}{|Z|} \angle (\delta_2 - \delta_1 + \theta_Z) \end{aligned} \quad (13.5)$$

To facilitate the subsequent analysis, we now assume a lossless line, with impedance $Z = R + jX \approx jX$. Note that the term “lossless” line generally refers to a line without resistance, where no real power is lost, from which it follows that $P_{12} = -P_{21}$. This does not give any information about Q_{12} and Q_{21} . In fact, there will be reactive I^2X “losses” on the line.

Substituting $\angle \theta_Z = 90^\circ$ and $|Z| = X$ helps separate the real and imaginary parts of the expression.

$$\begin{aligned} S_{12} &= \frac{|V_1|^2}{X} \angle 90^\circ - \frac{|V_1||V_2|}{X} \angle (90^\circ + \delta_1 - \delta_2) \\ -S_{21} &= \frac{|V_2|^2}{X} \angle 90^\circ + \frac{|V_1||V_2|}{X} \angle (90^\circ + \delta_2 - \delta_1) \end{aligned} \quad (13.6)$$

The voltage phase angle difference between the two buses, $\delta_1 - \delta_2$, is important and will be referred to often, so we'll abbreviate it δ_{12} .²² We use the equality $-\cos(90^\circ + \delta_{12}) = \cos(90^\circ - \delta_{12}) = \sin \delta_{12}$ to extract the real part of Eq. (13.6) from its second term, and account for the 90° with j before

²² We could also define $\delta_{21} = -\delta_{12}$, but that turns out to be unnecessary because $\sin \delta_{21} = -\sin \delta_{12}$ and $\cos \delta_{21} = \cos \delta_{12}$.

the imaginary part:

$$\begin{aligned} S_{12} &= P_{12} + jQ_{12} = \frac{|V_1||V_2|}{X} \sin \delta_{12} + j \left(\frac{|V_1|^2}{X} - \frac{|V_1||V_2|}{X} \cos \delta_{12} \right) \\ -S_{21} &= -P_{21} - jQ_{21} = \frac{|V_1||V_2|}{X} \sin \delta_{12} - j \left(\frac{|V_2|^2}{X} + \frac{|V_1||V_2|}{X} \cos \delta_{12} \right) \end{aligned} \quad (13.7)$$

The real part of Eq. (13.7) gives the expression for real power P_{12} that we already encountered as Eq. (7.1), and graphically in Figure 7.15. This important relationship is worth re-emphasizing:

$$P_{12} = \frac{|V_1||V_2|}{X} \sin \delta_{12} \quad (13.8)$$

Equation (13.8) holds as an exact equality only for lossless lines. However, it is often used as an approximation even for lines that are not exactly lossless, but are assumed to be dominated by inductive reactance. This relationship applies not just to transmission lines, but more broadly to power transferred across any device that can be approximated by a pure series reactance, such as the windings of a generator.

Equation (13.7) tell us that as the impedance (i.e., reactance) in the denominator increases, less power is transferred for a given choice of voltages. Conversely, holding impedance constant, we note that power transfer depends roughly on the square of the nominal operating voltage, since we expect both $|V_1|$ and $|V_2|$ to be within a few percent of this value. In the operational context, where line impedance and nominal voltage levels fixed, we are interested in the sensitivity of P_{12} and Q_{12} to the actual voltage magnitudes and angle difference.²³

From Eq. (13.8), it is clear the transmission of any real power across an inductive line necessarily requires some voltage phase angle difference. The greater the angle difference, the greater the real power flow—up to a point. This theoretical limit on the power that can physically be transferred is unrelated to the thermal rating of the transmission line. The theoretical maximum real power (for a lossless line) occurs at the peak of the curve in Figure 13.9 where $\delta_{12} = 90^\circ$, for which $\sin \delta_{12} = 1$. As discussed in Section 13.4.3, a power system cannot be operated in a steady-state anywhere near this theoretical limit, due to the risk of oscillations and ultimately losing synchronicity between generators.

Analogous to the $P-\delta_{12}$ relationship, there is a limit for power transfer associated with voltage magnitudes $|V_1|$ and $|V_2|$ as well as the line impedance. Such a limit applies to both real and reactive

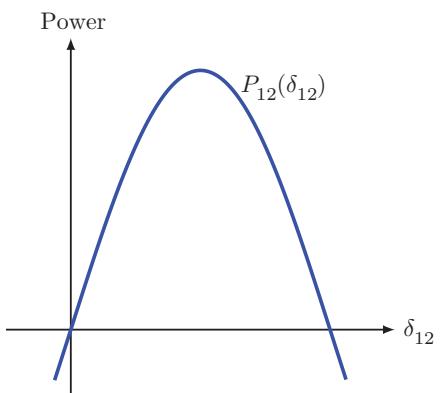


Figure 13.9 Real power transfer as a function of voltage phase angle difference across an inductive transmission line.

²³ Chapter 12 prominently features the finding that real power transfer is highly sensitive to the phase angle difference δ_{12} , while reactive power transfer has a stronger dependence on voltage magnitude, especially at the receiving end.

power. However, the explicit dependence of power transfer on sending and receiving end voltage magnitudes is not as easy to extract from Eq. (13.7) as the dependence on angle.

Complex power transfer depends on complex voltage, a two-dimensional quantity. To analyze how power varies with voltage phase angle difference, we assumed bus voltage magnitudes to be fixed. The curve in Figure 13.9 effectively takes a slice through a $|V|-\delta$ surface, to focus exclusively on the dependence of power on the angle δ while holding $|V|$ constant. This problem framing corresponds to the majority of practical situations.

Sometimes, though, it is important to explicitly examine the relationship between power transfer and bus voltage magnitudes. For example, reactive power resources such as capacitor banks, generators or static VAR compensators for regulating voltage at a certain bus may be unavailable or overtaxed, so that the bus voltage magnitude is no longer independently controllable. This situation mostly applies to radial transmission or distribution lines, with only load at the receiving end and no connection to other buses that actively regulate voltage.

In this case, as load increases, the unsupported bus voltage will tend to decline. Here we take a different slice through the surface, holding δ constant to focus instead on the dependence of power on $|V|$. Analogous to the relationship between power transfer and voltage angle, there will be a maximum theoretical transfer limit for real power relative to voltage magnitude. In practice, this condition is rarely reached—but if one attempted to increase the load beyond this maximum, the voltage would collapse, and the lights would suddenly go out. Below we analyze the (surprisingly nonobvious) mathematical relationship between power and voltage magnitude, for the case of a single radial transmission line.

First, we re-emphasize the *nonlinear* nature of the problem, where power demand is the independent variable and we are solving for an unknown bus voltage. This is very different from the problem where we are given sending and receiving end voltages for a line, and we wish to determine the amount of power flowing. When power is considered as a *function* of voltage, a unique solution for S_{12} and $-S_{21}$ can be unambiguously calculated from V_1 , V_2 , and Z using Eq. (13.5). But when we turn the equation inside out, it is no longer a function: there may be more than one voltage V_2 that satisfies some specified amount of power demanded (see the example in Section 12.4.5).

Mathematically, this makes sense because voltage is squared in the power equation, and a quadratic function can have zero, one, or two solutions. Conceptually, the ambiguity arises from the fact that we don't know the current or the lines losses. There will generally be two mathematically correct solutions for bus voltage, one of which corresponds to a realistic and the other to an excessively high current. In the high-current case, delivering the specified power to the receiving end bus is accomplished at the expense of huge line losses that drag down the voltage, in turn necessitating a higher current. This is the unstable situation to be strenuously avoided (Section 13.4.5).

Voltage varies with both real and reactive power demanded, and we can choose to construct either a $P-V$ or a $Q-V$ curve (often rotated and called $V-Q$). Again, we are slicing through a higher-dimensional space in order to express a relationship between two variables that can be depicted on a two-dimensional graph.

13.5.1 $P-V$ Curve

It is often reasonable to expect loads to have a relatively constant power factor, meaning that P and Q are tied fairly closely together. Accordingly, the relationship between real power and voltage can be conveyed by defining a parameter based on power factor, which is held constant for the purpose of drawing one two-dimensional curve at a time. With these general bearings in mind, let

us derive an expression for voltage magnitude in terms of real power delivered in a simple radial system (i.e., from a source across an impedance to a load, where the voltage magnitude at the load is not influenced by other sources from other directions).²⁴ The crux is to manipulate the equations so as to eliminate the voltage phase angle. The voltage phase angle difference δ_{12} will be implicitly determined but not explicitly stated in the context of studying the dependence of power on $|V|$.

To streamline notation, we drop the minus sign and define power demanded at the receiving end Bus 2 as $S_D \equiv -S_{21}$, where current is positive coming out of the transmission line. Also, we will simplify things by assuming the transmission line has only inductive reactance. This lets us start with Eq. (13.7) for the power received at Bus 2, which we call S_D and separate into real and imaginary parts for real and reactive power demanded:

$$\begin{aligned} -S_{21} &= S_D = P_D + jQ_D \\ P_D &= \frac{|V_1||V_2|}{X} \sin \delta_{12} \\ Q_D &= -\frac{|V_2|^2}{X} + \frac{|V_1||V_2|}{X} \cos \delta_{12} \end{aligned} \quad (13.9)$$

To get rid of δ_{12} , we use information about the load power factor. It will simplify notation and save space to define the parameter $\beta = \tan \theta$, a close cousin of the (displacement) power factor $p.f. = \cos \theta$.²⁵ Using β , we can write²⁶

$$S_D = P_D + jQ_D = P_D(1 + j\beta) \quad \text{or} \quad Q_D = \beta P_D$$

The algebraic trick to eliminating δ_{12} is to square Eq. (13.9), which allows us to use the identity $\sin^2 x + \cos^2 x = 1$. Before squaring the equation for Q_D , we rearrange it so that the cosine term is by itself on the right hand side:

$$P_D^2 = \left(\frac{|V_1||V_2|}{X} \right)^2 \sin^2 \delta_{12} \quad (13.10)$$

$$\left(Q_D + \frac{|V_2|^2}{X} \right)^2 = \left(\frac{|V_1||V_2|}{X} \right)^2 \cos^2 \delta_{12} \quad (13.11)$$

Then we substitute $Q_D = \beta P_D$ and $\cos^2 \delta_{12} = 1 - \sin^2 \delta_{12}$ to get

$$\left(\beta P_D + \frac{|V_2|^2}{X} \right)^2 = \left(\frac{|V_1||V_2|}{X} \right)^2 - P_D^2 \quad (13.12)$$

We now have an expression relating $|V_2|$ to P_D , where $|V_1|$, X and β are externally chosen parameters and δ_{12} does not appear at all. It is still quite messy to untangle.

A preferred arrangement is to write Eq. (13.12) as a standard quadratic equation of the form $ax^2 + bx + c = 0$ in the voltage-squared variable $|V_2|^2$:

$$|V_2|^4 + (2\beta P_D X - |V_1|^2) |V_2|^2 + (1 + \beta^2) P_D^2 X^2 = 0 \quad (13.13)$$

²⁴ This derivation format follows A.R. Bergen and V. Vittal, *Power Systems Analysis* (McGraw-Hill, 2nd edition, 2000).

²⁵ The angle θ is the difference between voltage and current phase angle. This β has nothing to do with other uses; there just aren't enough letters in the alphabet.

²⁶ Since $Q_D = S_D \sin \theta = P_D \sin \theta / \cos \theta = P_D \tan \theta$.

This format offers a tedious but reliable way to extract the two solutions to $|V_2|^2$ in terms of P_D and the given parameters.²⁷ The general quadratic formula²⁸ yields

$$|V_2|^2 = \frac{|V_1|^2}{2} - \beta P_D X \pm \left[\frac{|V_1|^4}{4} - P_D X (P_D X + \beta |V_1|^2) \right]^{1/2} \quad (13.14)$$

To most eyes, Eq. (13.14) does not immediately promote intuition. It is easiest to explore by substituting some representative per-unit values for the parameters.

For example, let the sending-end voltage be $|V_1| = 1.0$ p.u., the line reactance $X = 0.1$ p.u., and the power factor unity, making $\beta = 0$. With these values, Eq. (13.14) simplifies to:

$$|V_2|^2 = \frac{1}{2} \pm \left[\frac{1}{4} - 0.01 P_D^2 \right]^{1/2}$$

Here it is easy to see that at no load, when $P_D = 0$, the two possible solutions for $|V_2|$ are 1 and 0: either there is no voltage drop across the line (consistent with no current), or the voltage at the receiving end has collapsed. We can also see by inspection that a limit occurs at $P_D = 5$ p.u., which renders the expression inside the square root zero. This is the ultimate limit of the power transfer capability, since for any $P_D > 5$ p.u., there will be no real solution for $|V_2|$. Some useful intuition here is that in most realistic situations, we would encounter other constraints before ever reaching such a large power transfer. This is why voltage stability is not such a common topic.

Let's keep the same assumptions for $|V_1| = 1.0$ p.u. and $X = 0.1$ p.u., but introduce some reactive power demand at the receiving end, so that β becomes a positive number. For example, take $p.f. = 0.80$ lagging, so that $\theta = \cos^{-1} 0.8 = 36.9^\circ$ and $\beta = \tan 36.9^\circ = 0.75$. With these values, Eq. (13.14) becomes:

$$|V_2|^2 = \frac{1}{2} - 0.075 P_D \pm \left[\frac{1}{4} - 0.01 P_D^2 - 0.075 P_D \right]^{1/2}$$

By contrast, for a capacitive load with $p.f. = 0.90$ leading, we have $\beta = \tan -25.8^\circ = -0.48$, yielding:

$$|V_2|^2 = \frac{1}{2} + 0.048 P_D \pm \left[\frac{1}{4} - 0.01 P_D^2 + 0.048 P_D \right]^{1/2}$$

We can appreciate how the leading power factor will impart an upward sloping character to the curve, by way of the linear term $-\beta P_D X$.

Figure 13.10 shows the family of curves of $|V_2|$ versus P_D , plotted for three choices of $\beta = 0$, $\beta = 0.75$, and $\beta = -0.48$, using $|V_1| = 1.0$ p.u. and $X = 0.1$ p.u.

Let's walk through these curves for increasing power demand. When zero power is demanded, there will be no voltage drop. If we assume that the system is initially energized and not in any state of distress, it only makes sense for $|V_2| = |V_1| = 1.0$. As load demand $|P_D|$ increases (and Q_D along with it), the operating point moves along the curve. The voltage angle difference δ_{12} and the line current implicitly increase along the way, but these dimensions are not depicted here.

Initially, the curve is nearly horizontal, so the increase in power demand does not have much effect on voltage magnitude. This represents the typical, desired operating condition. With further increase in P_D , however, the unsupported bus voltage takes a turn for the worse: the downward slope steepens, and soon even a small additional increase in power drags the voltage down substantially.

²⁷ Note that there are really four solutions to voltage because the actual $|V_2|$ values could be negative, but this doesn't introduce any physical novelty; it just means we can reverse the polarity or reference direction.

²⁸ The quadratic equation $ax^2 + bx + c = 0$ has solutions $x = -b/2a \pm \sqrt{b^2 - 4ac}/2a$.

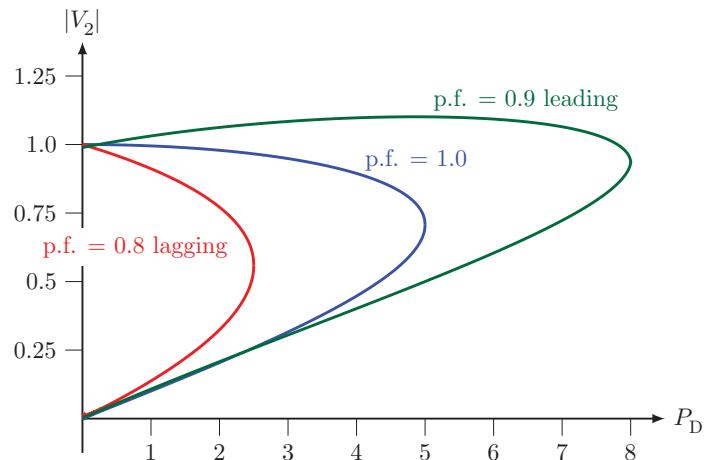


Figure 13.10 Family of P - V curves.

Though uncommon in practice, it is conceivable that the system could still be operating within its thermal limits, in which case no alarm bells go off as the operating point slides precipitously toward the power transfer limit known as the “nose” of the curve. This would be a point of no return, since the entire low-voltage side of the curve represents an unstable operating condition. Once on the wrong side, even control systems intended to shed load will be unable to drive the operating point back to the safe region, and voltage will inevitably collapse.

As for the role of power factor, it is consistent with earlier discussion (e.g., in Section 9.3.4): namely, that a lagging current will tend to result in a more pronounced voltage magnitude drop across an inductive line, while a leading current will tend to raise voltage magnitude at the load. This phenomenon is represented here by comparing the P - V graphs for the three different choices of β . With a lagging power factor, voltage declines much faster with increasing power, and the maximum real power transfer is significantly reduced. A leading power factor, by contrast, extends the P - V curve outward and upward, allowing for a voltage *rise* at the receiving end bus, before finally also curving downward at extremely high load. Even the capacitive load cannot ultimately get around the limits of power transfer capability—although it does push the threat of voltage instability out to higher current levels, which would more likely raise flags before the operating point approaches the nose.

Another important conclusion is that little can be done at the sending end to address a low voltage problem at the receiving end. While raising $|V_1|$ will shift the P - V curve upward, it does not change its basic shape. This is because no amount of generation or voltage control at the sending end (which would be accomplished with reactive power injection) can mitigate the voltage drop across the line. Accordingly, a common adage in power engineering holds that “voltage is local,” or “reactive power doesn’t travel.” In practice, a dangerously low voltage at any point in the network almost always means too much reactive power demand at that location. This reactive power must be supplied locally in order to restore the voltage magnitude to a safe level.

13.5.2 V - Q Curve

Finally, another slice through the multidimensional power-voltage relationship examines the mutual dependence between Q and $|V|$ at a given bus. The relationship can be described either as a Q - V curve (analogous to the P - V curve) or a V - Q curve (with Q on the vertical axis).

The simplest problem framing is for a radial transmission line serving a constant-power²⁹ load at the receiving end, as for the P - V curve. Here we refer to the imaginary part of Eq. (13.9),

$$Q_D = -\frac{|V_2|^2}{X} + \frac{|V_1||V_2|}{X} \cos \delta_{12} \quad (13.15)$$

where Q_D is the reactive power demanded at Bus 2, to be supplied by a voltage source at Bus 1.

To explore the dependence of Q_D on $|V_2|$, the most straightforward approach is to assign some constant values to the other variables, such as $|V_1| = 1.0$ p.u., and a fixed amount of real power with a corresponding value of δ . We can then plug in numerical values for a given situation. Because the control options for reactive power involve varying Q independently of P , we would not assume a constant power factor as we did for the P - V case.

To wrangle Eq. (13.15) analytically and eliminate the explicit reference to δ_{12} , one strategy is to employ the trigonometric identity $\cos^2 \delta + \sin^2 \delta = 1$ along with the real part of Eq. (13.9) to write

$$\cos \delta_{12} = \sqrt{1 - \sin^2 \delta_{12}} = \sqrt{1 - \left(\frac{P_D X}{|V_1||V_2|} \right)^2}$$

Substituting into Eq. (13.15), we obtain

$$Q_D = \frac{|V_1||V_2|}{X} \sqrt{1 - \left(\frac{P_D X}{|V_1||V_2|} \right)^2} - \frac{|V_2|^2}{X}$$

which simplifies to

$$Q_D = \sqrt{\left(\frac{|V_1||V_2|}{X} \right)^2 - P_D^2} - \frac{|V_2|^2}{X} \quad (13.16)$$

Figure 13.11 plots $-Q_D$ versus $|V_2|$ from Eq. (13.16) for a simple illustrative case with $|V_1| = 1.0$ p.u., $P_D = 0.5$ p.u., and $X = 0.1$ p.u. To resemble the most commonly seen presentation format, it displays reactive power *injection* rather than load at the bus, but this is an arbitrary choice since Q may be positive or negative in any case.

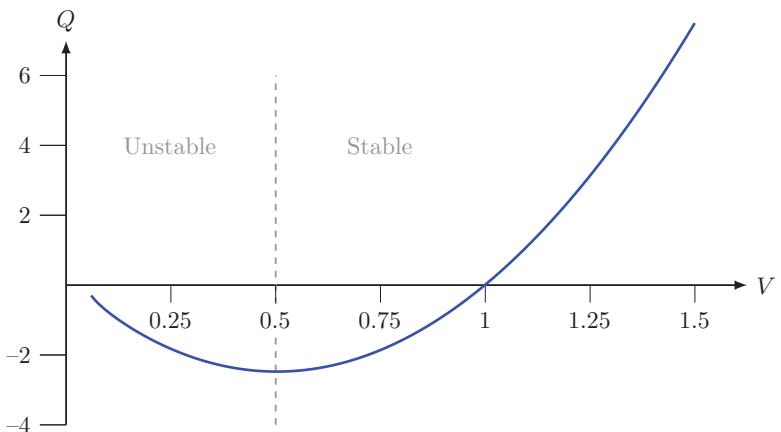


Figure 13.11 V - Q curve showing bus voltage magnitude versus reactive power injection (negative reactive load) for a simple radial system.

²⁹ Constant power here means that the load is deliberately specified, and does not change on its own with voltage.

The important take-away from Figure 13.11 is that in standard operation, we expect the curve to be upward sloping (such that an increasing reactive power injection should raise the local voltage)—but there is a *critical value* of Q , seen here as the bottom of the curve, which is analogous to the “nose” if we turned the graph sideways. This point corresponds to the minimum reactive power that needs to be injected at the given node (or conversely, the maximum reactive power that can be demanded by the load), at the minimum feasible voltage for that node. The point could lie above the x -axis and represent a minimum positive Q injection (i.e., requiring the load to be capacitive, or some VAR correction to occur there). The left side of the graph as displayed in Figure 13.11 represents an unstable condition. In reality, if one attempted to draw more (or inject less) reactive power than the critical value at this location, the voltage would collapse. Note that even on the right side of the curve, the slope declines as the operating point moves left, closer to the critical point: a small increase in reactive load will cause a pronounced reduction of voltage. This describes the *sensitivity* of voltage with respect to Q .

The vertical distance between the actual operating point and the critical value of Q is called the *reactive margin*. The horizontal distance between the actual operating point and the critical value of V is called the *voltage margin*. Crucially, in order to estimate reactive margin or voltage margin in the setting of a larger network, it is necessary to produce reference curves in an offline analysis. Like with the P - V curve, observation of the local voltage in itself does not indicate the distance to the cliff, since the curve depends on the operating state of the system as a whole.

In practice, Q - V or V - Q curves are produced through numerical simulations. For a real network, one would do this empirically by introducing a fictitious VAR source at a single bus of interest, model the VAR source as holding its bus voltage magnitude constant, and observe the necessary reactive power injection Q at this bus for different values of $|V|$ based on the power flow solution for the entire network.³⁰

It is common to produce a family of curves like in Figure 13.11 with different levels of real power demand P_D for each curve, based on a fixed network impedance. As P_D increases, the V - Q curve will move upward, indicating that the constraint on feasible values of Q at this bus is getting tighter. This reflects the fact that both real and reactive power demand at a given bus tend to draw the voltage magnitude down, and there is some trade-off between the two quantities.

Another important type of analysis holds load constant but experiments with different network contingencies, altering the value of X in Eq. (13.15). Recall from Section 13.3 that a secure operating condition means a network can tolerate a contingency, such as the sudden loss of a transmission line, without loss of load. Figure 13.1 illustrated a toy example where the concern is exceeding the thermal power flow limit on the remaining line, if a line parallel to it were to fail. In a related scenario, the concern would be that the loss of one transmission line effectively increases the impedance of the path, and thereby drives the system very suddenly into an unstable condition. Such an instability could involve either angle or voltage magnitude.

Contingency analyses in real power systems study scenarios like this by creating families of P - V , Q - V , or V - Q curves with impedance values that correspond to the loss of one or more transmission lines. Figure 13.12 shows a family of notional V - Q curves for a certain bus reflecting the loss of up to three transmission lines in the region. Suppose the initial operating point is on Curve A, at 0 MVAR. The loss of one transmission circuit would suddenly shift the system over to Curve B. Assuming no change in reactive power, this would push the bus voltage to the left. If a second line fails, the voltage drops further left, and perilously close to the critical point on Curve C. It is now imperative

³⁰ Prior to computer-based power flow analysis, engineers had to make their best estimates with pencil and slide rule. Unsurprisingly, standard design practice for many decades was to build systems with plenty of margin to accommodate uncertainty.

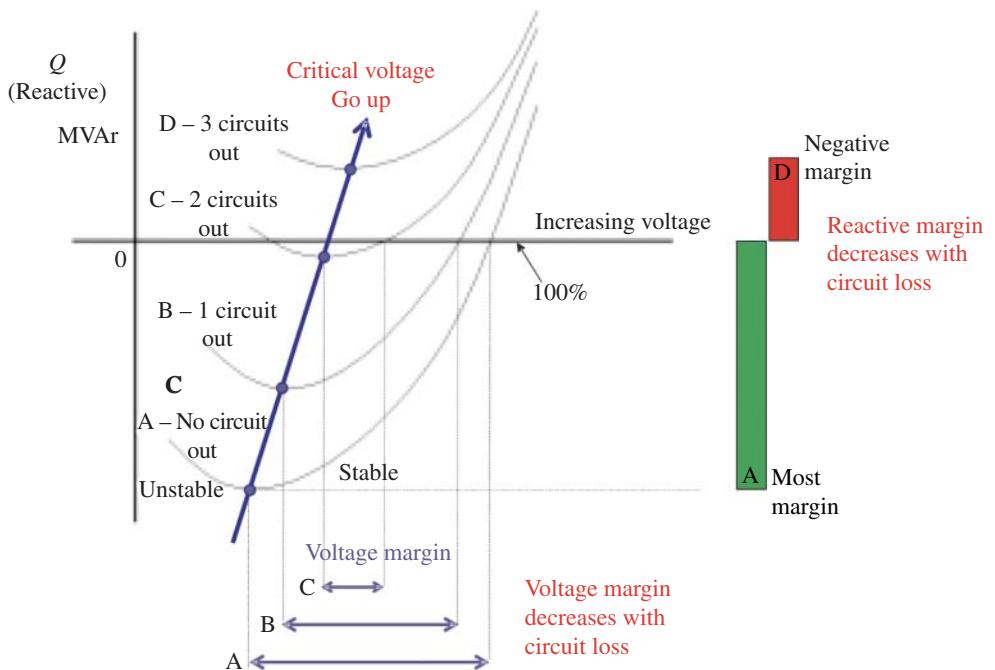


Figure 13.12 Family of V - Q curves for different transmission path impedances, illustrating the problem of voltage security. Source: Graphic from *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*, U.S.-Canada Power System Outage Task Force, 2004.

to recruit MVAR injection at this location, so as to move up and right on Curve C. Else, the third contingency will put the system below Curve D, with no physically feasible operating point, and the lights go out due to voltage collapse.³¹

Problems and Questions

- 13.1** A 60-Hz, 4-pole steam turbine generating unit is rated 500 MVA, 12 kV and has inertia constant $H = 4 \text{ p.u.-sec}$.
- What are the synchronous electrical and mechanical angular frequencies, ω_s and $\omega_{m,s}$?
 - Find the rotational inertia constant J in $\text{kg}\cdot\text{m}^2$.
 - What is the unit's rotational kinetic energy at synchronous speed, in joules?
 - What are the electrical and mechanical angular accelerations, α and α_m , if the unit is at synchronous speed and the steam turbine is supplying 500 MW of mechanical power, but zero power is delivered to the grid?
 - Suppose the unit is initially operating at $P_m = P_e = 0.8 \text{ p.u.}$, with $\omega(0) = \omega_s$ and $\delta(0) = 10^\circ$, when a fault reduces generator electrical output to zero. Assuming the

³¹ As mentioned earlier, voltage collapse is not, in fact, how the lights went out on August 14, 2003. The proximate cause of the blackout was the tripping of protective relays due to low voltage and high current, after the loss of several transmission lines. Operators' lack of awareness of shrinking reactive and voltage margins, which would have offered an indication of the system's insecure state overall, kept them from taking timely corrective action.

mechanical power remains constant and ignoring damping, calculate the power angle δ after 2, 4, 6, 8, and 10 cycles from onset of the fault.

- 13.2** Graph the power angle versus time from part (e) of Problem 13.1. Does it increase linearly? Why or why not? Explain.
- 13.3** Suppose that in the fault scenario in Problem 13.1, the normal connection between the generator and the grid is restored after four cycles, at which instant the electrical power delivered jumps to the value corresponding to $P_e(\delta)$. Assume the generator and its normal connection path to an infinite bus on the grid is characterized by an impedance of $Z = j0.217 \text{ p.u.}$, and that sending- and receiving-end voltages are both approximately 1.0 p.u.
- Find δ and $P_e(\delta)$ after four cycles of faulted operation.
 - Find ω_m after 4 cycles of faulted operation.
 - Do you expect the unit will swing back to normal? Explain. (Assume there is some damping available, but you don't need it for your calculation.)
- 13.4** For the fault scenario in Problem 13.1 (a complete loss of power output P_e to the electric grid, while the turbine power P_m remains unchanged) use the equal area criterion to estimate the maximum time in cycles that can elapse in the faulted condition before the generator would be unable to recover when the normal power balance is restored, and state the maximum value of δ that it reaches during such an excursion.
- 13.5** Consider the following modifications to Problem 13.1:
- Suppose the operating frequency is changed to 50 Hz, while the physical generator stays the same. How would the values of $\omega_{m,s}$, H , and rotational kinetic energy at synchronous speed change?
 - Suppose the frequency remains at 60 Hz, but the number of poles is changed from 4 to 8. How would the values of $\omega_{m,s}$, H and rotational kinetic energy at synchronous speed change?
 - Qualitatively, do you expect either of these two variations (50 Hz or 8 poles) would affect the ability of the unit to sustain a severe power imbalance?
- 13.6** A lossless transmission line has inductive reactance $X = 0.06 \text{ p.u.}$ The sending end voltage is $|V_1| = 1.0 \text{ p.u.}$
- Suppose the power demand at the receiving end is $P_2 = 3.0 \text{ p.u.}$ at a power factor of 0.90 lagging. Calculate $|V_2|$ using Eq. (13.14).
 - Determine the voltage angle, current, and reactive losses. Is this a stable operating condition?
 - Suppose that the transmission line, which consists of two parallel circuits, experiences a fault on one circuit, so that the impedance suddenly doubles to $X = 0.12 \text{ p.u.}$ What do you expect happens? Explain.
 - Repeat the above exercise with power demand $P_2 = 3.0 \text{ p.u.}$ at unity power factor.
 - For the unity power factor case, find the maximum value of P_2 that determines the "nose" of the P - V curve for $X = 0.06$ and 0.12 p.u.

14

Power Electronics

This chapter offers a brief introduction to the basic physical principles of power electronic devices and some of their most important applications in power systems. The main objective here is to appreciate the profound differences between modern solid-state technology and the large electromagnetic devices on which the legacy grid was built. At the time of this writing, a vast transition is in progress where power electronics are replacing traditional loads and rotating generators, while also appearing in new functions to control electric power. But because of the sheer scale of the legacy investment, existing assets can't just all be replaced overnight. A necessary part of innovation—at least in places where electric grids have existed for many decades—is the effective integration of old and new. This requires some appreciation for the very different capabilities and constraints of different technologies that are expected to operate together.

Many readers, especially engineering students, will already be much more familiar with modern power electronics than with old-fashioned power systems. This chapter aims to provide some general context for this knowledge and remind younger readers of what wasn't obvious to previous generations. The chapter is also intended to give readers from nonengineering disciplines some elementary insights and intuition.¹

14.1 Power Conversion: Introduction

It is often desirable or convenient to convert electric power from one form into another. "Form" here has two dimensions: voltage magnitude, and time signature. By time signature we mean whether voltage and current are direct or alternating (d.c. or a.c.), or the frequency of the alternating current, but it could also mean just a shift in the phase (i.e., the exact time of the zero crossing) of the waveform.

The most classic example of a power converter is the transformer, a completely analog physical device that works only with alternating current. Of course, transformers are so important to power systems that they have their own chapter earlier in the text. A transformer converts voltage magnitude while doing nothing to change the a.c. frequency.² In the days of Edison and Westinghouse, transformers were the only practical technology for stepping voltage up or down at the power system scale, which in turn motivated the use of a.c. for commercial systems.

1 For a more detailed and rigorous treatment, see for example Ned Mohan, *A First Course in Power Electronics* (Wiley, 2011) or R.W. Erickson and D. Maksimovic, *Fundamentals of Power Electronics* (Springer, 3rd edition, 2020).

2 A closer look reveals that a transformer will in fact slightly alter the waveform, but this is due to imperfections such as magnetizing currents or saturation in the core, and not part of the transformer's job. While it may introduce harmonics, this does not affect the fundamental 50- or 60-Hz frequency.

Changing d.c. voltages is much harder. A transformer faced with a d.c. current will do nothing but heat up. Some 19th-century options included using different numbers of batteries in series, or placing different numbers of loads in series. This scheme has the grave disadvantage that loads connected in series cannot be operated independently.

Another option is to place two rotating machines back-to-back on the same shaft, one motoring and the other generating, with their electricity in a different form (voltage and/or frequency). This is called a *rotary converter*, which enabled conversion between a.c. and d.c., along with different voltages. For example, a d.c. motor can power an a.c. generator. It works, but it's clumsy: there are moving parts that take up space, need to be maintained, and incur losses. Besides the inefficiencies, rotary converters had poor frequency regulation because a d.c. motor's rotational speed is sensitive to the input voltage. This technology has been completely superseded by power electronics, but some of the most basic principles are the same.

Changing the time signature of electric power always requires opening and closing some electrical connection, at a suitable time interval. In a rotary converter, this job is done by the brushes of the d.c. motor or generator, which alternately contact opposite ends of the machine windings. In this sense, the classical, analog process of power conversion is based on a rotating mechanical switch. The big revolution for power converters came with the ability to perform electrical switching in a different physical medium, which would not only increase the efficiency and ease of operation, but allow for the on-off switching cycle to be vastly sped up. Moreover, the switching process could be logically isolated from the power flow itself, allowing for an external signal to control the on-off state at any moment within the cycle.

Besides the ability to make and break an electrical connection at will, a power conversion device requires two other physical features: the ability to impose a directional preference for current flow and the ability to store some energy for a fraction of a cycle. Like switching, these functions have been increasingly miniaturized as power electronic technology evolved.

14.2 Legacy Power Conversion Technologies

14.2.1 Mercury Arc Valves

Well before the advent of semiconductors, mercury vapor arose as an important medium for making and breaking electrical connections. The mercury arc valve³ became the go-to component anywhere that called for conversion to or from high-voltage d.c., until the 1960s. Some of these arc valves are still in operation today, in the 2020s.

The basic operating principle of the mercury arc valve is that mercury vapor becomes electrically conducting in the presence of free electrons and an electric field, by sustaining an arc of ionized mercury atoms and electrons. The unique property of mercury is that unlike other electrical conductors, it is liquid at room temperature and readily vaporized. When a negative voltage is applied to a pool of liquid mercury, its surface, acting as the *cathode*, emits electrons. These electrons fly across the interior space of the valve—an evacuated glass bulb, containing mercury vapor at low pressure—toward a positive terminal, the *anode*. Along the electrons' path, mercury atoms lose and regain electrons, emitting ultraviolet and visible light as electrons settle back into atomic orbitals. The resulting arc is analogous to lightning through air, but much more controlled. Positively charged mercury vapor ions float toward the negative cathode.

³ The device is more graphically described by its German name, *Quecksilberdampfgleichrichter*, which translates literally as “quicksilver vapor equal router.” This is most certainly the author’s favorite word in her native language.

The anode is typically made of carbon, which conducts but does not emit many electrons if heated or negatively charged. If the polarity is reversed (with the positive charge on the pool of mercury and the negative charge on the carbon), nothing much happens. Consequently, the arc valve acts as a *rectifier*, or combination of an on-off switch and a directional element: when presented with an alternating voltage, it conducts in one direction, but blocks current in the other. The reservoir of liquid mercury continually replenishes its surface and the vapor, so that the cathode does not wear out even under large currents over long duration. Multiphase arrangements have multiple anodes, where the arc moves from one to another during the cycle.

A device that is simply nonconducting during half the cycle is called a *half-wave rectifier*. For a more continuous output, it can be paired in a circuit with another converter of opposite polarity to make a *full-wave rectifier*, where the d.c. output of constant polarity is alternately supplied by different circuit branches. The principle, illustrated in Figure 14.1, applies regardless of the physical nature or medium of the converter.

Suppose an alternating voltage is supplied on the three-phase system at the top of the illustration. (In keeping with German convention, the phases are labeled RST instead of ABC, and K stands for *Kathode*.) The squiggly lines in the middle represent transformer windings. When anode A_1 is positive, it attracts electrons and draws an arc from K. Thus, while K serves as the source of negative charge to the arc valve, it appears to the d.c. circuit below as a sink for electrons, or a positive terminal. When the voltage diminishes at A_1 and increases at A_2 , the arc relocates. During the half cycle when the voltage at A_1 is negative, no current flows in phase R. In the full-wave version on the right, either anode 1 or 2 will always have a positive polarity relative to the center tap on the transformer, which is connected to the negative d.c. terminal. This provides for a continuous load on each a.c. phase.

One additional design modification allows active control of the current flow, by introducing a *grading electrode* in the middle of the valve. The application of a small voltage to this electrode can prevent an unintentional reverse current by repelling electrons, and can also adjust the timing of the

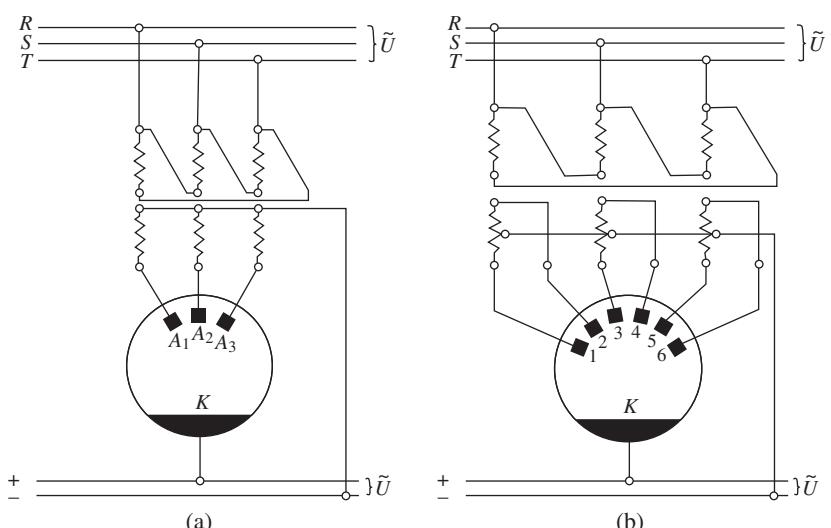


Figure 14.1 Mercury arc valve, configured as (a) a half-wave (three anodes) and (b) full-wave (six anodes) rectifier. Source: Wdwd-Own work/Wikipedia/ CC BY 3.0.

forward current flow. This basic concept is analogous to the control of triodes and transistors, discussed further below. Mercury arc valves of this design can sustain voltages on the order of 100 kV and up to hundreds of amperes of current, making them practical for bulk power transmission.

14.2.2 Vacuum Diodes and Triodes

The same principle of one-directional conduction applies to vacuum tubes with *thermionic emission*. These devices are smaller, with solid electrodes that cannot sustain the high current levels of a liquid mercury cathode. When the solid cathode is heated, it emits electrons—named “cathode rays” by their German discoverers in the late 1800s, before the electron had been identified—which fly through the vacuum and are attracted to the anode or *plate* on the opposite side of the tube.⁴ Since only the cathode is heated (and only electrons, not these positive ions, can fly), a *diode* (with two terminals) conducts current only in one direction. An ideal diode presents an infinite resistance in the reverse direction. A realistic diode will have some nonlinear voltage–current characteristic, of the sort discussed further below in the context of semiconductors.

As the name suggests, a *triode* has three electrodes: a cathode, a plate, and a *grid* that serves the function of controlling the electric field across the vacuum and thereby adjusting the conductivity. Figure 14.2 shows a basic schematic.

Though tangential to power systems, the historical importance of this development cannot be overstated: a small voltage signal applied to the grid of a triode could now be used to control, by direct analog relationship, a much larger current from cathode to anode. Specifically, increasing and decreasing the voltage applied to the grid at specific times establishes or stops the flow of electrons from the cathode, by attracting or repelling them. This opens the option of using a much stronger power source, which in itself is not easily modulated, to supply the current to the device. Thus was born *amplification*, and with it the entire era of electronics, in the early 1900s. From telegraph and radio transmission to rock & roll powered by tube amps, the ability to control an electric current with an independent small signal came to define a century. The next step for this amazing new functionality was to make it ever smaller.

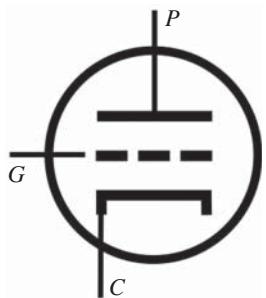


Figure 14.2 Basic schematic of a vacuum triode, with labels for cathode, plate (anode), and grid.

14.3 Solid-State Technology

Solid-state technology brought circuit switching to the microscopic level. As the name suggests, everything happens in the solid phase, with no vacuum, vapor or plasma required. The key, of course, lies in *semiconductors*: materials whose electrical conductivity can vary depending on the specific conditions applied to them. Miniaturizing an electric switch also introduced the ability to replicate and interconnect vast numbers of switches, to change their on-off status in a coordinated manner: in other words, computing.

The basic building block of a semiconductor circuit is the *p–n junction*, a sharp boundary between layers of material with different properties, which allows an electric current to flow in one direction

4 The same basic physics underlies *cathode ray tubes* (CRTs) in television screens and computer monitors predating the 1990s. In the CRT screen, these electrons land on phosphor dots that visibly light up, their path deflected by a precisely applied electric field to create an image.

but not the other. Stacking another p–n junction on top creates a switch, where the application of a voltage or current to the middle layer makes the entire sandwich conducting or non-conducting. Such a device is called a *transistor*.

The first working transistor dates back to 1947, its significance recognized by the 1956 Nobel Prize in physics. Today, there are many different types of transistors. It almost seems preposterous to attempt a textbook section on this vast subject, but we'll scratch the surface just enough for a most basic appreciation of what happens inside these devices, and what they can do for us in the power systems context.

14.3.1 p–n Junctions and Diodes

The p and n in the p–n junction stand for positive and negative charge carriers that prevail inside the p-type and n-type semiconductor material, respectively. For simplicity, let's assume the material is silicon; the operating principles are the same for other semiconductors. Like carbon, silicon (atomic number 14) has four valence electrons, which are available to make bonds with neighboring atoms. Also like carbon, silicon atoms arrange themselves into a regular crystal analogous to diamond, in which every atom is connected to four neighbors in covalent bonds (i.e., bonds made by two electrons equally committed to each of the two atoms, which represents a stable quantum mechanical state that physically locks the atoms together).⁵ In this pure form, the silicon crystal does not conduct electricity, because the electrons are bound and not free to move.

But realistically, at the atomic scale, things are always on the move. Just from thermal energy, a small fraction of electrons randomly abandon their bonds, leaving behind a vacancy or *hole* (and a lonely former partner electron) where a bond would be expected.⁶ Electrons from adjacent bonds can then snap over to fill in, propagating the location of the vacancy. Thus, electrons and holes equally appear to travel around in the material.

Each silicon atom in the crystal contains a number of protons in the nucleus equal to the number of electrons surrounding it, so that there is no net electric charge. When an electron abandons its bond, it carries a negative charge to a place where this is not balanced. Conversely, the hole it leaves behind appears as a net positive charge. These holes, or absences of electrons, can be treated as if they were positive objects carrying charge as they travel through the material. In this sense, a current can be conducted either by positive or negative charge carriers traveling through the crystal.

A pure semiconductor contains an equal number of positive and negative charge carriers. Their prevalence compared to the overall number of atoms is tiny, because the electrons need to be excited with extra energy—the *bandgap*—to leave their bound state.⁷ We now introduce *doping*, or small impurities that alter the balance. The dopant consists of an element that occasionally takes silicon's place in the crystal lattice, which has a number of protons and valence electrons different from silicon. Electrically, the dopant atom is still balanced, but it creates a systematic excess of free electrons or holes in the lattice structure. For example, the dopant may be phosphorus (atomic number 15) with five valence electrons, or boron (number 5) with three valence electrons. But a fifth valence electron has no neighbor in the crystal to bond with, and with only three valence electrons, one place for a bond remains empty. The presence of excess electrons or holes in the otherwise regular crystal, even though it does not introduce any net charge, creates the ability to transport current.

⁵ Silicon is abundant in the earth's crust, and it forms crystals much more easily than diamond, which requires very high temperature and pressure.

⁶ In photovoltaic (PV) cells, large numbers of electron-hole pairs are created by incident sunlight.

⁷ For example, in silicon at room temperature, there are on the order of 10^{10} free electrons and holes (*intrinsic carriers*) per cubic centimeter. This sounds like a lot, until we recall Avogadro's number. There are on the order of 10^{22} Si atoms per cm^3 , so only about one in a trillion atoms gives up an electron to join the *conduction band*.

Specifically, phosphorus-doped silicon with its excess electrons has an abundance of negative charge carriers and is therefore called an *n-type material*. Even though occasional holes exist, they are orders of magnitude fewer in number, and electrons are called the *majority carrier* in such a material. A current consisting of extra holes—the *minority carrier* here—would quickly vanish, because any new holes simply *recombine* with electrons. Conversely, boron-doped silicon is called a *p-type material* because its holes act as the majority, positive charge carriers.⁸ It is important to remember that p-type and n-type materials in themselves are still electrically neutral, but either type of doped material will conduct electric current.

Interesting things happen at the boundary or *junction* between a p-type and n-type material. Essentially, it is a dance between electric charge and entropy: the force of electrical repulsion among like charges, versus the force of randomness that drives electrons and holes to *diffuse* across the junction. Any free carriers near the junction are likely to recombine—that is, electrons will settle into the gap created by a boron atom, while traveling holes will vanish in the vicinity of a phosphorus atom, filled by its extra electron. This recombination makes the crystal bonds look tidy, but gives rise to a charge imbalance: the region doped with boron now acquires a net negative charge (since its protons do not compensate for all four electrons surrounding it). The region doped with phosphorus acquires a net positive charge (since it needed five electrons, not four, to compensate for its positive charge).

This local accumulation of net charge is eventually limited by electrical repulsion. But in a narrow region near the p–n junction, the diffusion of charge carriers dominates over the electrical force. With the two forces in balance, a permanent *intrinsic electric field* is established across the boundary.⁹ In order to create a decisive intrinsic field, the p–n junction must be very abrupt. Sophisticated manufacturing processes allow for the extremely precise diffusion of dopant atoms into the crystal, to form boundaries only a few atoms wide.

When using the p–n junction as a directional element or *diode* to control externally applied currents, the key property is that the *space charge region* near the boundary, while carrying a local charge imbalance, is effectively depleted of free electrons and holes that would be available to travel and conduct current. This is called the *depletion zone*, illustrated in Figure 14.3. Within the depletion zone, the bonds throughout the crystal lattice look tidy (with two electrons each), but beneath this order lies a charge imbalance due to the different number of protons of the dopant atoms.

Suppose now that we want to drive a substantial current across the junction, by externally applying a voltage (referred to as *bias*). The direction is critical. The “forward” direction is that which counteracts the intrinsic electric field, supplying a positive voltage on the p-type side. The excess holes on the p-type side neutralize the space charge region and refill the depletion zone with available carriers. Similarly, excess electrons populate the depletion zone on the n-type side. Consequently, there are now available charge carriers throughout, and the p–n junction conducts in the forward direction.

In the reverse direction, however, the external voltage adds to the intrinsic electric field. In doing so, it adds more electrons to the p-type material, which fills holes and thereby extends the depletion zone; similarly, the external voltage pulls electrons away on the n-type side and again adds to the space charge region and widens the depletion zone. With a much wider zone devoid of

⁸ The concentration of majority carrier is on the order of a million times that of the intrinsic carriers.

⁹ In a PV cell, this field has the important property of driving any electrons and holes, knocked loose from their bonds by incident sunlight, in one direction—thus creating a voltage at the terminals of the cell. The current generated by sunlight in this way flows in the reverse direction of what the diode conducts when supplied with external voltage.

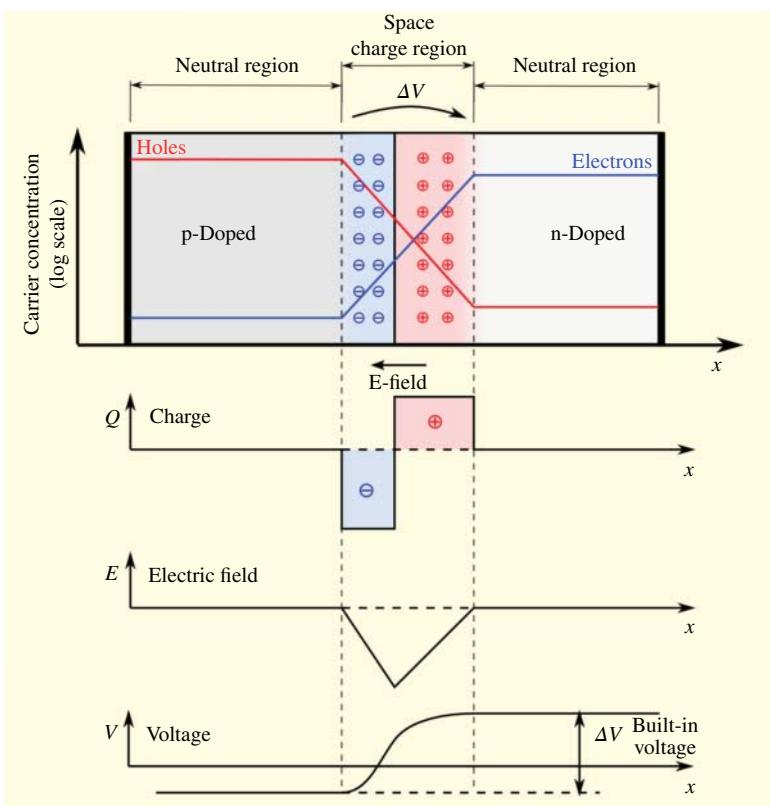


Figure 14.3 p–n Junction and electric field. Source: TheNoise/Wikipedia/CC BY SA.3.0.

carriers, the p–n junction is no longer conducting, and thus blocks current in the reverse direction. In this way, a microscopic one-directional circuit element or *diode* is created. A diode will block current under reverse bias unless the external voltage is large enough to overwhelm the insulation presented by the depletion zone. This process, which may or may not be reversible, is called *avalanche breakdown*. An *avalanche diode* is specifically designed to become conducting in the reverse direction at the *breakdown voltage*, and return to its nonconducting state when the external voltage is reduced.

14.3.2 Transistors

A transistor is a piece of semiconductor material that changes from a conducting to a nonconducting state in response to a small voltage or current signal, making it act like an electrical switch. Because the transition happens on the atomic level and involves no mechanical movement, it is not only very compact but extremely fast, allowing circuits to be switched back and forth many thousands of times per second. This process lies at the heart of information technology that represents data in terms of open and closed circuits (0s and 1s).

A basic type of transistor called a *bipolar junction* transistor (BJT) is created by adding a third layer of semiconductor to a p–n junction to form a sort of sandwich. (The term “bipolar” indicates that the device recruits both types of charge carriers, electrons and holes.) The layers can be either n–p–n

or p–n–p.¹⁰ With an external voltage applied in either direction, one of the two junctions inside will be forward and the other reverse biased, making the whole device non-conducting by default.

Crucially, a third external connection point to the center layer, called the *base*, can activate the transistor into a conducting state. This could work in either direction according to the external bias applied, but practical transistors are made for use with d.c. and designed asymmetrically for efficiency. For a BJT, the *emitter* side has a higher concentration of dopants than the *collector* side, which reduces the amount of carrier recombination loss.

When a voltage is applied to the base, a current flows through the middle layer and through the “easy” way (across the forward biased junction). However, this populates the entire middle region with more charge carriers. For example, in an n–p–n transistor, a positive voltage at the gate will inject holes and draw electrons into the p region. The presence of these carriers in the middle region is enough to counteract the depletion of the reverse-biased junction, causing it to conduct as well. The amount of current conducted between emitter and collector transistor can be many times greater than the current injected at the gate; this ratio is known as *gain*.

Depending on the connection in an external circuit and the voltage bias applied to it, a transistor may serve as an amplifier (for either voltage or current or both), or as a switch, where it toggles between fully “on” and fully “off” positions by varying the base current. Because of the gain, a very small electrical signal applied to the base can control a much larger power flow, or trigger the transistor to switch on and off.

An alternate design called the *field effect transistor* (FET), such as the metal oxide semiconductor type (MOSFET), uses a more subtle mechanism for flipping the switch. Instead of injecting a current to flood the central region with charge carriers, a mere voltage applied at the terminal called the *gate* acts to modify the depleted region or *channel* and make it conducting, by way of creating an electric field across it.¹¹ By using voltage rather than current as the trigger, FETs can be made smaller and dissipate less power; they can also have higher switching frequencies.¹² One practical constraint of field effect transistors is that they do not tolerate very high voltages.

The best of all worlds—speed, efficiency, robustness, and actuation by voltage rather than current—can be had with the most recently developed type, the *insulated gate bipolar transistor* (IGBT).¹³ The insulated gate implies that no current is injected here, but an applied electric field suffices to trigger the IGBT into its conducting state. The majority of modern power electronic devices used for power conversion in the electric grid context are built on IGBTs. These include modern solar inverters, power conditioning devices such as solid-state transformers, and advanced power converter stations for high-voltage d.c. lines. Typical switching frequencies for this purpose are on the order of tens of kilohertz.

A key advantage of the IGBT is that it can be deliberately turned “on” and “off” by applying a positive or slightly negative voltage to the gate, to become conducting or nonconducting. Such devices are categorized as *voltage-source converters* or VSCs. The significance of this characterization for power systems becomes clear only in contrast to the legacy alternative, thyristors, which are *current-source converters* (CSCs).

¹⁰ These work the same way but with opposite current polarity, which is sometimes convenient to combine in circuit design.

¹¹ A field effect transistor is a *monopolar* device, as only one type of charge carrier (electrons or holes) contributes to the current flow through the channel (p or n). Analogous to emitter, base and collector, the FET terminals are called *source*, *gate* and *drain*.

¹² There are vastly more subtleties to the many types of transistors used in different circuits, and the functional limitations have a lot to do with the particular application—for example, whether the purpose is amplification of signals, or we are simply storing information.

¹³ The three terminals of the IGBT are called *emitter*, *gate*, and *collector*.

14.3.3 Thyristors

The *thyristor* is an important device that for many decades dominated high-voltage a.c.–d.c. conversion technology, in applications such as high-voltage d.c. connections underwater or between asynchronous a.c. grids. It is synonymous with *silicon-controlled rectifier* (SCR). Thyristors are current-controlled, which imposes significant practical limitations on thyristor-based converter stations. We include this topic here because of the large number of existing thyristor-based assets in the field, which are not about to be retired overnight. In summary, their limitations compared to VSCs are that (i) current-source converters cannot independently create an a.c. waveform signal, relying instead on an already established alternation of current on their a.c. side, and (ii) they cannot control power factor, but always consume reactive power. The most important practical implication is that energy delivered to an a.c. grid through thyristor-based converters cannot be used for a *black start* to initially energize an a.c. grid when recovering from an emergency event. What follows is just enough physical detail to offer some support for the above statements.

The key physical features of a thyristor are that it may only conduct current in one direction and can be actively switched on, but not off. Unlike the BJT, which stops conducting as soon as the base current vanishes, the thyristor *latches* in its “on” state after receiving only a brief current pulse at the terminal called the *gate* (analogous to the base), and remains on as long as a positive load current flows from *anode* to *cathode* (analogous to collector and emitter). If placed in a circuit with an alternating current, the thyristor remains conducting until the load current reaches zero, and remains nonconducting until triggered again with a base current.

The rationale for this behavior is not immediately obvious. Thyristors are built out of four layers of differently doped semiconductor, with three junctions inside. In the classic p–n–p–n arrangement, the gate is at the middle p layer. With a positive external bias applied (positive at p and negative at n), which is the forward biased direction for the whole device, two of the three junctions always conduct. Its middle junction is in the n–p orientation relative to that bias, however, so it is reverse biased and generally nonconducting. With an opposite external voltage, there are two reverse biased junctions, so the thyristor definitely wants to stay off.

By applying a positive voltage at the gate, some current flows easily toward the cathode (across the junction that was forward biased anyway). The presence of these charge carriers also helps fill the depletion zone for the middle reverse biased junction, and makes it conducting. Since the first junction was already forward biased, the entire thyristor is now in its “on” state. So far it works just like an n–p–n transistor that is turned “on.”

The important difference is that a thyristor is designed in such a way (by choosing a suitable combination of internal voltages and currents) that the central junction experiences an *avalanche breakdown*, where the presence of some (sufficiently energetic) charge carriers triggers the creation of additional charge carrier pairs at that junction. This avalanche means the junction remains conducting even after the gate current trigger goes away. In a typical transistor such a design would be counterproductive, because the device would just always stay on. But the additional p layer of the thyristor gives it a directionality, and will put a stop to the conduction if and when the direction of current flow tries to reverse. Consequently, it is only the load current (also referred to as *line current*) that determines when the thyristor reverts to its “off” state. This is why it is called a current-source converter or CSC.

Because a thyristor only conducts in one direction, it acts as a primitive *rectifier*. In this very crude manner, a single thyristor can convert alternating current into an intermittent current of a one-directional polarity.

The thyristor is controlled by a small voltage pulse delivered to the gate with very precise timing, which results in a small current that triggers the “on” state for forward current flow. The device remains conducting throughout the cycle until the moment that the current reaches zero, then strictly blocks any reverse current. It remains off until another gate trigger pulse occurs in the next cycle, under forward current. The timing is accomplished with an external circuit that detects the zero crossing of the a.c. voltage and has an adjustable delay for sending the pulse.

By adjusting the timing of the pulse, the duration of the thyristor’s “on” state for each a.c. cycle is controlled. This timing is expressed in terms of a *firing angle* α in degrees (relative to the full 360° cycle). The effective magnitude of the d.c. output voltage—which in turn determines the power transferred by the thyristor—is given by the average over a cycle. Thus, the power can be continuously controlled through α .

One key observation is that for turning the thyristor “off,” the direction of current, not voltage, matters. Note that these are not synonymous when the thyristor finds itself in a circuit with a power factor different from unity. In the common case of a slightly inductive circuit with a lagging current, the zero crossing of the current occurs slightly later in the cycle than the zero crossing of the alternating voltage.

This leads to the somewhat counterintuitive situation where the thyristor remains “on” while the current is still positive, even though the voltage has already reversed. Recall that instantaneous power always equals the product of instantaneous voltage and current. This tells us that for the fraction of the cycle where voltage and current are opposed but the thyristor is still “on,” the power flow is negative. The average power transferred over the course of each cycle is completely analogous to the derivation of average power in Chapter 3.

A single thyristor will be “off” for at least half of each cycle. By arranging multiple thyristors in a circuit such as a *full-bridge rectifier*, with opposite orientation relative to the a.c. source, and firing them in precise coordination with each other, they will become alternately engaged over the course of a cycle so that the circuit remains conducting for both halves of the cycle.¹⁴ This produces a much smoother voltage and current output, even though the waveform is still choppy-looking.

The choppiness is synonymous with a waveform having high harmonic content (see Section 5.3). In a practical circuit, this would be remedied by introducing an inductor as a low-pass filter. Ignoring the details of the waveform, we observe the fundamental component of the current appears lagging relative to the a.c. voltage by the firing angle α . Both of these factors—the presence of an inductive filter, and the delay of current due to α —contribute to the thyristor appearing as a consumer of reactive power.

If there is a power source on the d.c. side that continually sustains the current in the face of a negative voltage, such a thyristor rectifier bridge can act as an inverter. In this case, α is timed between 90° and 180° so as to keep each thyristor “on” primarily during the opposite part of the cycle, where the a.c. voltage is negative although the current (driven by the d.c. source) is still positive. Consequently, power is transferred from the d.c. to the a.c. side.

Crucially, while the thyristor inverter can impart energy to the electrons on the a.c. side and thus force a current flow opposing the direction of voltage, it cannot itself produce the alternating voltage signal. If the a.c. side is not already energized with an alternating voltage, the power transfer simply won’t work. Therefore, thyristor-based inverters or CSCs are useless for attempting a black start after a grid outage. Also, a thyristor has no way to regulate timing of current independent of the average power transferred, and therefore cannot control power factor.

¹⁴ This coordination aims to avoid gaps in time between either directional circuit branch conducting, but also and especially overlaps where opposing branches would be “on” simultaneously, causing a short circuit known as *commutation failure*.

These drawbacks notwithstanding, thyristors are remarkable devices in that they can withstand very high voltages and currents, for countless cycles without significant degradation. The main failure mode, mentioned in a footnote above, is *commutation failure*, wherein one thyristor in a bridge unintentionally remains “on” for a moment after its counterpart has also been switched “on,” causing a short circuit. This can happen because of some distortion of the waveform, where the current just doesn’t go to zero as expected. Again, the vulnerability of the thyristor stems from the inability to deliberately turn it “off” at a specific instant in the cycle.

14.4 Inverters

Inverters are a very important subset of power converters. Generally, an *inverter* is a device that changes direct current (d.c.) into alternating current (a.c.), meaning that it has to change the current’s direction, or *invert* it, many times per second to produce the desired a.c. frequency. Inverters are used wherever there is a d.c. electric source, such as a battery, PV module or fuel cell, but a.c. power is needed to either serve a specific appliance or inject power into the grid.

Because d.c. sources are often at a low voltage, inverters may also include a transformer component to produce the desired a.c. output voltage. Applications include stand-alone PV systems for remote facilities, homes, and vehicles as well as small- and large-scale power generation within the electric grid (see Section 15.2). Inverters also make up one aspect of converter stations for high-voltage d.c. (HVDC) transmission, a classic application of the thyristor-based converters described in Section 14.3.3. Finally, inverters are used in the process of converting a.c. from one frequency to another—for example, the *wild a.c.* output from a variable-speed wind turbine to a steady 50 or 60 Hz.

To say that inverter technology has evolved significantly over the past several decades, even the past few years, is an understatement. Historically, interfacing between a.c. and d.c. was something of an exotic topic that might be skipped entirely in an introductory power systems course. Well into the 1980s, most solar PV systems were stand-alone installations. With the rapid growth of grid-connected solar applications around the 1990s came a new set of concerns about power quality as delivered by inverters, prompting new performance standards. More recently, in the 2000s and 2010s, a broad range of control capabilities have been overlaid on the basic inverter function of approximating a sine wave by rapidly throwing tiny switches.

14.4.1 Basic Inverter Function

As of this writing, a wide range of inverter models are on the market that employ different circuit designs for inversion and power conditioning, specifically built to interface with the legacy grid. Many of these have advanced capabilities to transfer power in a meticulously controlled manner, allowing them to help regulate voltage and frequency in the network.

In Section 14.3.3. we already encountered the important distinction between voltage- and current-source converters (VSCs and CSCs). Only a voltage-source converter can act as a *grid forming* inverter, since its job is to create a waveform at the desired frequency from scratch. Such a device can also be operated in *grid-following* mode, where it will inject current at specific times relative to the given a.c. voltage. Modern inverters in microgrids can transition between grid-forming and grid-following modes of operation, depending on whether they are connected to the main grid, or disconnected and creating their own power island. The vast majority of inverters in modern power applications such as renewable energy are built on VSCs.

The many differences among inverter designs involve size, reliability, tolerance with respect to varying input, and efficiency. Modern units in the kilowatt and megawatt range operate at efficiencies in the high 90s of percent. These sophisticated devices are a world apart from, say, the kind of cheap inverter that plugs into a car's 12-V d.c. outlet to produce something resembling a 120-V a.c. source, at tens of percent of losses that make it hot to the touch.

A key criterion for inverter performance is the *waveform*, or how closely the a.c. output approximates a mathematically ideal sinusoidal wave. Early inverters that appeared commercially in the 1970s simply reversed the direction of voltage and current 120 times per second, creating a *square wave* of 60-Hz frequency whose magnitude could be easily stepped up with a transformer. Such a square voltage waveform can create big problems due to unwanted heat. As detailed in Section 5.3.2, waveform distortion is quantified in terms of *harmonic content* or *total harmonic distortion* (THD). The harmonics, which essentially make up the "corners" of the wave, do not contribute usable energy to the load at the fundamental frequency. For example, in a motor, the harmonic components cannot create torque, but instead act to overheat motor windings; in electronic and audio equipment, they might create buzzing.

A slight and still inexpensive improvement was offered by the *modified square wave*, sometimes called a *quasi-sine wave* shown in Figure 14.4. It can be produced in essentially the same way, using a switch and a transformer, except that the voltage and current are kept at zero for a brief moment between every reversal. Another technique is to first increase the d.c. voltage with a step-up converter (a device that creates a high voltage by assaulting an inductor coil with extremely fast-changing current obtained from rapid switching) and then chopping the output; this makes for reduced weight and increased efficiency. In either case, the width of the nonzero voltage pulse is made such that the area under the curve, which is proportional to the amount of power transmitted, is the same as it would be for a sinusoidal wave of equal amplitude. This makes the modified sine wave more compatible with motor loads. The harmonic content is still high, but not as high as for a square wave.

With more elaborate switching schemes, the modification can be carried further to create what may be called a *modified sine wave*, or simply referred to as a "sine wave inverter." Here the cycle is divided into many smaller segments, creating a stepped voltage function that resembles a sine

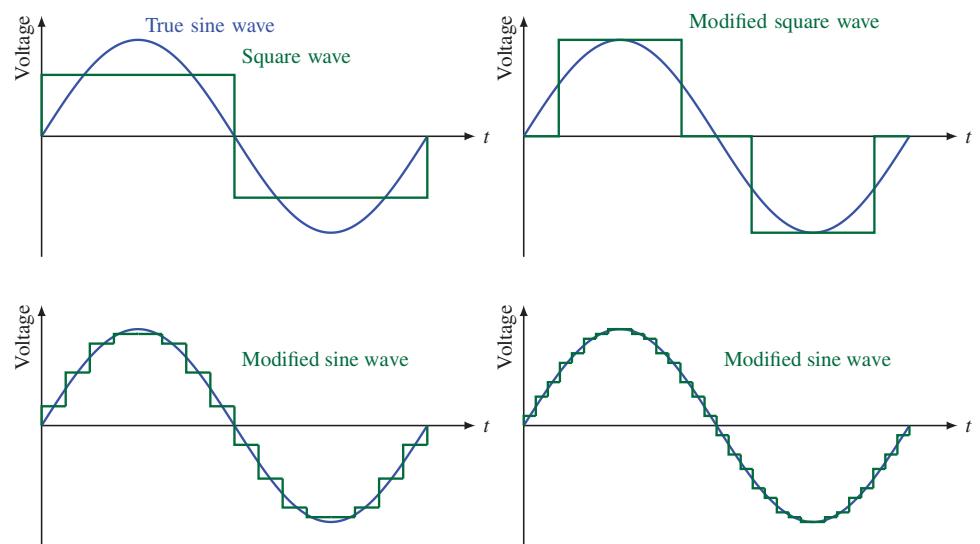


Figure 14.4 Inverter waveforms.

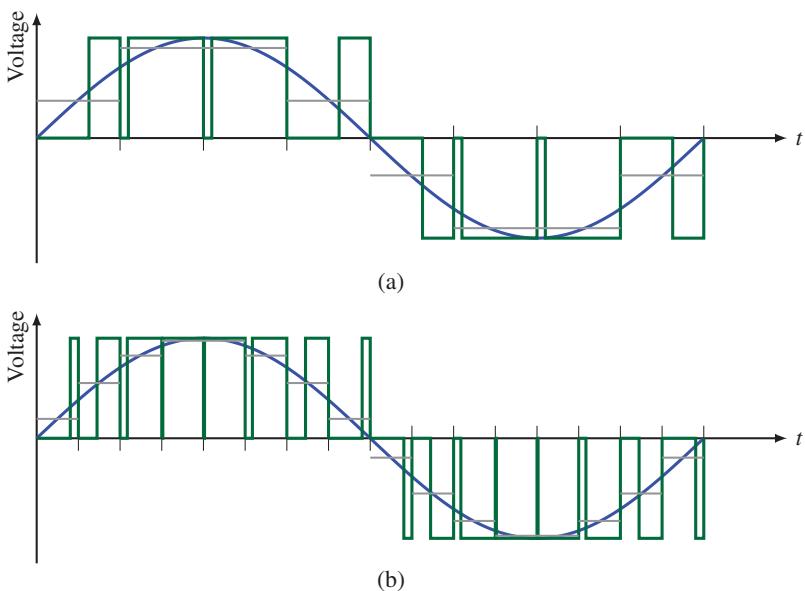


Figure 14.5 Pulse-width modulation with 8 cycles (a) and 16 cycles (b). Horizontal gray lines represent the average value of each pulse.

wave much more closely. For most practical loads, the output is indistinguishable from an ideal sine wave.

A modern approach for generating a “true” sine wave is called *pulse-width modulation* (PWM). This involves switching the voltage on and off very rapidly, at up to tens of kilohertz, in intervals that get longer and shorter throughout the a.c. cycle, so as to create voltage and current pulses of varying width. Figure 14.5 illustrates the concept with a small number of pulses for visual clarity. From the standpoint of power transfer, the changing duration of the pulse (called the *duty cycle*) has the same effect as a varying voltage magnitude, because we can think of the voltage and current as being averaged over a small portion of the a.c. cycle. Thus, during the portion of the cycle where the pulses are wide, the circuit effectively sees the full voltage amplitude; when the pulses are narrow, they appear to the circuit as a diminished voltage. In that sense, PWM effectively represents a signal of changing magnitude.

Another approach to creating a sine wave by rapid switching involves adding a set of transformers with different turns ratios whose outputs can be combined to yield different voltage magnitudes. For example, there are 27 possible combinations for the sum of three voltage terms that can each be either positive, negative, or zero. By comparison, modulating the pulse width can produce the equivalent of several hundred steps per cycle, which is therefore smoother. The transformer approach also makes for a heavier unit (due to the weight of the transformer cores) but can have high efficiency.

Note that the illustrations in Figures 14.4 and 14.5 are simplified for clarity and show fewer steps than an actual inverter unit would produce. The output will likely also be “cleaned up” after the switching process by *filtering* the harmonics. This is accomplished simply by placing an inductor in series, which presents a higher impedance to the higher frequencies (Section 3.3.1). The switching frequency in PWM is so rapid that the harmonics associated with the corners of each pulse are easily filtered by a comparatively small inductance.

14.4.2 Sample Inverter Circuit

Having discussed the function, how do we actually create an inverter circuit? The basic idea in designing any power converter is to combine switches (transistors), directional elements (diodes) and energy storage elements (capacitors and/or inductors) into a circuit *topology* that collectively will result in a voltage and current of the desired magnitude and time signature. Analyzing the output of a particular topology involves walking through an a.c. cycle in intervals during which voltages and currents at various points in the circuit increase, decrease, or remain constant.

The simplest example to illustrate a relevant circuit topology for our purposes is a full-bridge inverter, shown in Figure 14.6. V_{DC} represents a constant, d.c. voltage source, and the inverter circuit is intended to produce an alternating output voltage V_{AC} across the load. Switches S_1 , S_2 , S_3 , and S_4 are toggled on and off by an external signal, whose timing is of the essence.

In the case of a CSC using thyristors as switches (Section 14.3.3), a gate pulse triggers the ON state, and the reversal of current on the output side by external means (owing to the connection to an active a.c. grid) is required to initiate the OFF state. For a VSC using MOSFETs or IGBTs (Section 14.3.2), an external voltage signal turns the transistor ON and OFF, with switching frequencies that can be orders of magnitude greater than the a.c. waveform. Figure 14.6 uses the graphic symbol for an IGBT, whose triggering signal is received at the open circle on the left hand side. Shown in parallel on the right side of each switch is a *flyback diode* whose purpose we will address in a moment.

The switches are opened and closed in diagonal pairs, with two states for the circuit. When S_1 and S_2 are ON and S_3 and S_4 are OFF, the output voltage V_{AC} has a positive polarity, aligned with V_{DC} in the reference direction from top to bottom. Current flows through S_1 and S_2 from top to bottom, and the circuit branches with S_3 and S_4 conduct zero current, so can be temporarily deleted from the diagram. All four diodes are oriented with their forward direction pointing upward, so that with the applied voltage they are *reverse biased* (Section 14.3.1) and thus nonconducting.

After a suitable time interval has elapsed, the connectivity of the circuit is reversed, so that S_1 and S_2 are OFF and S_3 and S_4 are ON. We may now delete the branches containing S_1 and S_2 , and by tracing the closed circuit find that the load is connected to the source V_{DC} with S_3 and S_4 on either side. Crucially, though, the top of V_{DC} now feeds the bottom of V_{AC} , causing its polarity to reverse, along with the direction of current flowing through the load. After another time interval, the circuit returns to the previous state, again reversing the polarity. Thus, it has converted V_{DC}

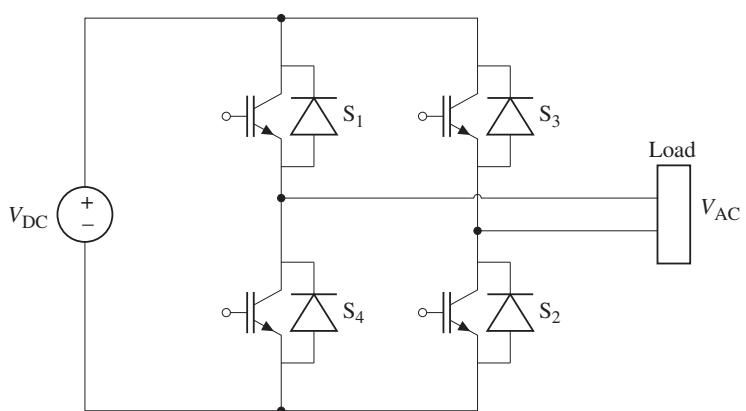


Figure 14.6 Full-bridge inverter circuit.

into an alternating voltage as presented to the load. To generate a square wave, the circuit would simply be switched each half-cycle, at 100 or 120 Hz.

If the load were purely resistive, our design could be complete. However, loads in general may be somewhat inductive or capacitive, which means that the timing of the current they draw will be lagging or leading relative to an alternating voltage. For example, if the load is inductive and its current is lagging the voltage, some current will want to continue to flow in the positive direction even after the voltage has turned negative.¹⁵ Immediately after switches S_1 and S_2 have turned OFF, this lagging current will flow back to the source (entering at the positive side and thus acting to recharge the source) through the flyback diodes at switches S_3 and S_4 , which are oriented to conduct in this direction (while the diodes at S_1 and S_2 are still reverse biased and do nothing). After some delay depending on the load power factor, the lagging current will cease and finally reverse, now flowing through the on switches S_3 and S_4 in alignment with the source voltage. When S_3 and S_4 turn OFF, the pattern repeats with the flyback diodes at S_1 and S_2 providing the return path.

An alternative topology to the full-bridge shown here is the *half-bridge* inverter. This is simpler in that it involves only two alternating switches, but harder to absorb visually because the source voltage is cut in half by a pair of capacitors, and the load connected to the halfway point. The concept of alternating the polarity is the same.

For reasons discussed above, a simple square waveform is undesirable. One option would be just to add a series inductance to filter the harmonics. Alternatively, the full-bridge topology shown is compatible with the PWM approach. Note that PWM cannot be done with thyristors but requires VSCs because switches must be turned off independently and much faster than the alternating current on the load side reverses direction.

The PWM signal is generated by a *comparator* circuit that is left off standard diagrams because it is a well-understood component that uses very little power.¹⁶ The comparator is a logic circuit that determines the precise moment in time to issue switch opening and closing commands, using a low-voltage signal (and essentially zero current) to the transistor gate. The circuit compares the voltages of two inputs and produces opening and closing commands at the instant when one input is greater than the other, and vice versa. Figure 14.7 shows how the desired PWM pattern, with pulses that become wider as the effective a.c. voltage magnitude increases, is produced by the intersection of a sine wave of the desired voltage output shape called the *modulating signal*, and a triangular wave called the *carrier wave* at the nominal switching frequency. Unlike the overall inverter output, the wave signal for this purpose is easy to obtain from a very small analog source or *oscillator* because it need not transfer power.

14.4.3 Inverter Control

The switching of an inverter circuit as described in Section 14.4.2 is buried deep on circuit boards within the device. When speaking of inverter control, we are usually referring to the inverter's interaction with the external load circuit, where actions are orders of magnitude slower than the pulsed switching that generates the shape of the sine wave output. We distinguish *inner* and *outer control loops*, with inner control loops being several times faster. To use a biological analogy, PWM is like nerve cells firing; the inner control loop is like a heartbeat; and outer control is like walking.

¹⁵ The fact that the current is now flowing "uphill" against the voltage means that the load is expending energy and doing work on the source during this portion of the cycle, as detailed in Section 3.4.2.

¹⁶ Recall from Section 14.2.2 that the revolutionary property of the transistor is to separate the control functionality from the power transfer itself.

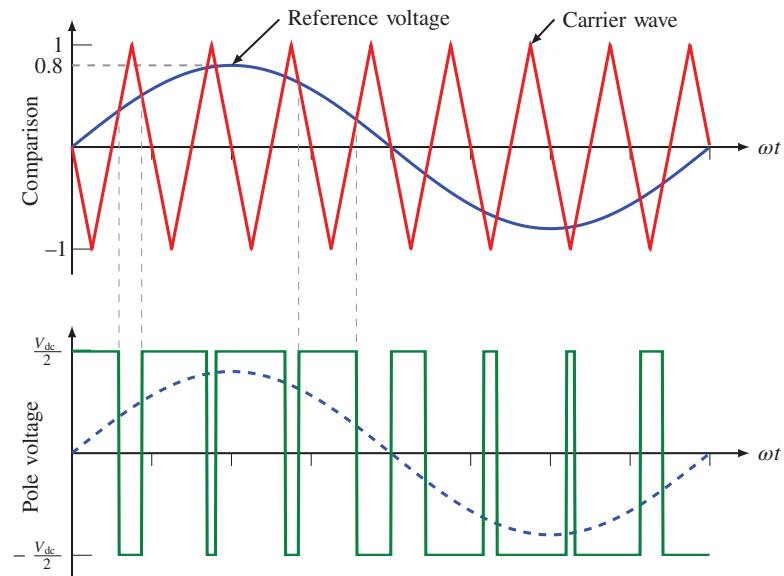


Figure 14.7 PWM signal, simplified for clarity. Actual carrier wave frequency would be much greater.

For example, in grid-following mode, inner control consists primarily of a *phase-lock loop* (PLL) that matches the timing of the voltage waveform to the external a.c. circuit. It provides a signal to the PWM modulator to adjust its phase and duty cycle based on direct measurements of the inverter output voltage and current. The outer control loop is concerned with actions such as adjusting the real and reactive output power relative to the voltage measured on the external circuit. These latter types of control actions are the focus of this section. Note that commercial inverters make even the outer control loops quite opaque to the user, who may be left with a few high-level settings to select—like choosing one of several destinations to walk to.¹⁷

Depending on the control algorithms (or lack thereof), inverter-based generation in power systems can represent either a liability or an asset to the grid at large. Particularly when associated with variable solar and wind power, inverter-based resources can appear volatile and fickle—especially to the grid operator accustomed to the traditional steam and hydroelectric generators around which power systems were designed. But the flip side of lacking the inherent inertia and predictability of large, rotating machines is being unconstrained by their physical limitations. Inverters have a lightning-fast response time, and their control instructions can be as creative as desired. When coupled with just a modicum of energy storage or headroom in the prime mover capacity, the instantaneous power injection from inverters can be modulated at will. When this capability is harnessed to compensate for variations, disturbances or oscillations from other sources, inverters have the potential to act as a valuable resource for grid balance and stability.

As of this writing, many advanced inverter functionalities are in various stages of research and development. One example is *synthetic inertia*, in which the inverter-based resource mimics the desirable aspects of the behavior of a rotating machine—namely, to limit the systemwide rate of change of frequency in the face of sudden, large generation-load imbalances. Another example is using the inverters at HVDC converter stations (Section 7.2.4) to compensate for

¹⁷ This opacity makes sense for several reasons including safety, liability, and protecting intellectual property. However, it presents challenges for microgrid operators wanting to coordinate equipment from different manufacturers, not to mention researchers wanting to test novel control strategies with off-the-shelf equipment.

wide-area subsynchronous oscillations (i.e., slower than the a.c. frequency, typically less than 1 Hz), analogous to the working principle of noise-canceling headphones.

Even common off-the-shelf inverters are capable of performing many functions whose details are beyond the scope of this chapter. Internal to a solar PV system, a key inverter feature is *maximum power point tracking* (MPPT), which makes the most of the available solar power by adjusting the impedance presented to the PV module or array. This allows the d.c. operating voltage to vary, while maximizing the product of voltage and current as the sunshine level varies, and keeping the output a.c. voltage steady.

Facing the grid, important inverter functionalities relate to power quality, including voltage and power factor corrections, and algorithms to protect both the equipment from damage and the distribution circuit from being accidentally energized during an outage. These capabilities, along with specific performance criteria and validation procedures, are detailed in the IEEE Standard 1547, which has seen profound revisions since the early 2000s. Inverters that interconnect with the electric grid to supply power will generally be required to comply with this standard, while the local utility may request specific control features to be enabled or not. Some of the most important inverter functions specified in IEEE 1547 include anti-islanding protection, low-voltage and low-frequency ride-through, and volt-VAR and volt-watt control.

The purpose of anti-islanding protection is to ensure that some local portion of the grid is not energized by one or more inverters after an outage elsewhere in the system to cause an *unintentional power island* (see Section 7.1.6). Unlike microgrids that are designed to operate independently in an intentional manner, with controllers that match generation and load, a cluster of inverters and loads that accidentally continues to operate without any connection to the main grid would not be prepared to ensure a steady frequency and voltage, potentially damaging customer equipment. Worse, if the status of an energized section of the distribution system is not known to utility personnel, this poses a serious hazard during repair or restoration after an outage. Meanwhile, the fact that the power went out in the first place suggests a heightened chance that crews would come to the area, and that there is some problem (say, tree contact or a downed line) that could be hazardous if energized before it is repaired.

Recognizing an islanded state is not trivial. How would an inverter know? In principle, a power outage on the grid should be recognizable by the fact that the voltage goes to zero, and there is no more a.c. frequency to follow. In reality, grid events span a large gray area of voltage and frequency disturbances. At high penetration levels of distributed solar generation, when there are many other inverters in the neighborhood, the voltage may not quickly decay when power from the substation has been cut off, even if no inverter has been assigned the grid-forming task. Therefore, the voltage magnitude and frequency could still bear some resemblance to the nominal, expected values. The question is, how far of a voltage and/or frequency deviation is reliable evidence that the local grid is islanded?

Historically, it was considered safest to err on the side of assuming that any significant departure from nominal voltage and frequency indicates a power island, and that inverters should therefore trip offline within a short time when certain narrow voltage or frequency limits are exceeded. On the other hand, it is not desirable to instruct inverters to disconnect during most other grid disturbances. Especially with a growing amount of inverter-based resources on the grid, more damage might be done by inverters overreacting and exacerbating a disturbance by tripping offline, than by staying online for a while to energize an unintentional island. For example, if the system frequency drops due to a sudden loss of generation and many inverter-based generation units trip offline in response, this would be exactly the wrong thing to do, and could risk taking

the whole grid down.¹⁸ Instead, inverters should “ride through” a range of voltage and frequency disturbances, detailed in the more recent IEEE 1547 updates.

The graphic representations of these requirements show either voltage or frequency on the vertical axis and time on the horizontal axis, somewhat analogous to the ITIC curve in Figure 5.1 in that more extreme departures from nominal conditions are tolerable for shorter durations. The standard distinguishes areas where grid-connected inverters must trip, where they must not trip, and where they may trip at their local discretion. Still, there can be ambiguity. For instance, consider that “frequency” is not a well-defined property of a signal that is not, in fact, periodic (see Section 16.2.4). Thus, drawing the line between an islanding event and other grid disturbances remains a difficult decision, and any rule or algorithm is almost guaranteed to occasionally get it wrong—especially considering that the dynamic behaviors of the entire system are evolving.

Another innovation that has come with increased penetration levels of inverters on the grid is first the permission and then the requirement to actively assist in regulating voltage. This applies to inverters of all sizes, but has been a most challenging issue for smaller-scale generation such as rooftop solar. Historically, it was seen as introducing too much uncertainty, and potential adverse interactions with legacy voltage regulation equipment, for distributed resources to take part in managing the voltage profile on a distribution feeder. Thus, it was common to operate inverters on whatever real power output they had, with a fixed power factor regardless of voltage.

By analogy, the situation is a bit like telling the children to stay out of the kitchen while dinner is being prepared. Yet as inverter-based resources collectively make a significant power contribution relative to the load, and thus have a significant effect on distribution feeder voltage, it may become not only economical but necessary to recruit them for active voltage management. Fortunately, the kids have grown up: power electronic devices are eminently capable, programmable, and obedient. The crux is knowing what to tell them to do, so that they can most helpful.

Generally, we expect that more power injection at a particular location will tend to raise the voltage level. This could be either real or reactive power. Depending on the power factor of the overall load on the circuit, and depending on the X/R ratio of inductive reactance to resistance of the distribution line, either real or reactive inverter output may be more effective or advantageous for helping regulate voltage magnitude (some light on this issue is shed by the LinDistFlow equations in Section 12.6).

Since inductive behavior tends to dominate on lines as well as loads, voltage is usually more sensitive to reactive power, and volt-VAR control is most common. This also makes sense because the solar resource cannot be arbitrarily increased unless there is also energy storage, or some amount of real power output is held back to provide headroom for up-regulating generation. Whether watts or VARs, the idea is to increase power injection when voltage is low, and to dial it back when voltage is too high.

The IEEE 1547 standard specifies volt-VAR and volt-watt *droop curves* that indicate a deadband of normal voltage during which the inverter should operate at unity power factor, and a guideline for increasing power injection at low voltage or decreasing it at high voltage. These curves are piecewise linear, with inflection points that can be adjusted as a matter of local policy. Figure 14.8 illustrates a sample volt-VAR droop curve with hypothetical values.

Since a modern inverter can easily operate at either a lagging or leading power factor, that is, “generate” or “take in” VARs, the range of reactive power contribution spans $\pm 100\%$ of its capacity.

¹⁸ Such an event occurred in 2016, where well over a gigawatt of solar generation in California disconnected (as per programmed instructions) in response to the electrical transient caused by a transmission line failure, in the Blue Cut Fire incident. See https://www.nerc.com/pa/rrm/ea/1200_MW_Fault_Induced_Solar_Photovoltaic_Resource/_1200_MW_Fault_Induced_Solar_Photovoltaic_Resource INTERRUPTION_Final.pdf (accessed March 2023).

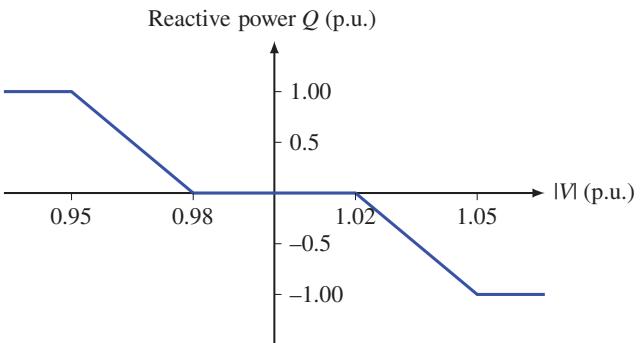


Figure 14.8 Sample volt-VAR droop curve. The per-unit (p.u.) scales indicate a decimal fraction of nominal voltage and maximum available power, respectively. The corner points that define the curve are specified based on local utility requirements.

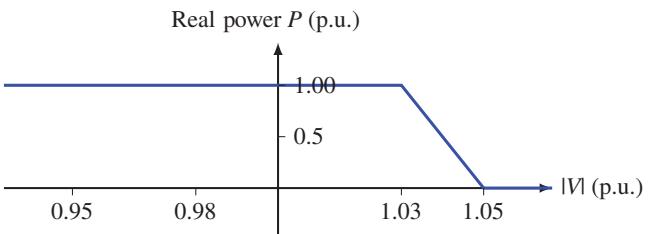


Figure 14.9 Sample volt-watt droop curve. The corner points that define the curve are specified based on local utility requirements.

Analogous to the reactive capability curve of a synchronous machine (Figure 10.23), an inverter has a capacity limit based on total current or apparent power (kVA). Therefore, a positive or negative VAR contribution may come at the expense of the ability to generate (and sell) real power if the inverter is operating at its limit: there is no energy cost, but potentially an opportunity cost. The ability to inject or consume both real and reactive power is known as *four-quadrant operation*, which can be visualized as an extension of Figure 10.23 to the left (where taking in watts requires some load or battery). Figure 14.9 illustrates a sample volt-watt curve that could be used for curtailing real power output during overvoltage conditions. In the example shown, the resource is not expected to absorb real power, but simply to stay within the area underneath and left of the curve.

To date, interconnection tariffs for small-scale inverter-based resources in the U.S. typically do not include voltage support or other grid services; in fact, they may prohibit active voltage control. As policies related to renewable and distributed electricity sources continue to evolve, their active participation in managing electric grid conditions—along with suitable economic incentives—may become increasingly sophisticated, as well as routine.

14.5 FACTS

Another important use of power electronics in power systems is the direct modification of the network properties with the aid of solid-state technology—essentially, transistors scaled up and combined to handle large power applications—in a category of hardware called *flexible a.c. transmission systems* (FACTS).

Transmission lines are generally assumed to have fixed physical parameters such as length and impedance that become firm constraints in modeling and analysis. Other components such as transformers and capacitors may have variable states or settings, but conventionally these settings are discrete and require mechanical switching. FACTS technology offers ways to modify the electrical characteristics of transmission components much more rapidly, even in real time, so as to increase operating efficiency and relieve constraints without the need for adding major hardware. FACTS devices include various types of reactive compensation, phase shifting, and power flow control. The idea is to effectively change the impedance of a given transmission link as seen by the system on an instantaneous basis, by means of an appropriately designed solid-state electronic circuit.

For example, a long transmission line may have a stability limit less than its thermal limit, meaning that no more than a certain amount of real power can be transmitted along it without risking a loss of synchronicity between the generators at either end. Conventionally, this limit poses a firm constraint on permissible power flow scenarios. For instance, it could mean that generators in location *A* must not inject more than a certain amount of power based on generation at *B*, lest the flow from *A* to *B* (specifically, the voltage phase angle difference between *A* and *B*) becomes too great. The only way around the constraint would be to upgrade the link with a larger or additional conductor that provides a lower impedance. FACTS technology, by contrast, allows system operators to intervene directly by modifying the behavior of the transmission link in question: in this case, a solid-state device would shift the voltage phase angle on the line between *A* and *B* in such a way as to reduce the difference, despite the generators' difference in production. We could say that FACTS amounts to cheating the power flow calculation from Chapter 12 by actively controlling flow along a particular link.

This should seem to contradict the basic principles emphasized throughout this text: that current and power flows throughout a network are determined solely by boundary conditions such as generation, load, and impedance, and are therefore beyond our immediate control. But the key to achieving control lies in the time dimension. In a conventional electric power circuit, we assume the properties of the hardware components, including the connections among them, to be unchanging. Solid-state technology, however, affords the ability to switch circuits around many times within a single a.c. cycle. This means being able to strategically open and close connections between elements such as transformers, inductors, or capacitors that add or subtract a.c. waveforms at specific points during the cycle.

The ability to exercise local control over the voltage waveform has many possible applications. In the phase-shifting example above, we alter the timing or phase angle of the voltage. This can be done on a continuous basis, to shift the steady-state angle separation across a transmission link, or it could be done dynamically, for example, to counteract oscillations in the system. We can also change the shape of the voltage waveform to eliminate unwanted harmonics or resonance; or we can control the voltage magnitude, or its relationship to current, to control reactive power flow.

Thus, by effectively entering a different time dimension, fast-acting power converters like FACTS sidestep the conventional rules of analog a.c. power systems. The practical limits for this technology mainly reside in the relationship of capital cost to operational gains for any particular installation. One likely opportunity for the economical implementation of highly controllable power converters on a large scale is in conjunction with high-voltage direct-current (HVDC) transmission, as discussed in Section 7.2.4.

15

Resources

15.1 Generation Resources

Many texts on power systems begin with a chapter about energy use and electric generation resources to lay out the broader significance of the power engineering subject. This book assumes that most readers already have some familiarity with the energy sector, motivating their study of the electric grid. Accordingly, this chapter aims to serve as a simple reference on the major types of power generation and energy storage and their most interesting attributes for our context (where the length of each subsection is not proportional to the contribution of each resource). We emphasize newer technologies that are increasingly important in the process of decarbonizing electricity, or shifting away from fossil fuels.¹

Since the beginning of the electric power industry in the late 1800s, there have been essentially two types of power generation facility: hydroelectric plants, which convert the downhill movement of water into rotation of an electric generator, and thermal plants, where the turbine-generator is pushed by the force of hot, expanding steam. Both types use synchronous generators (Section 10.3), with some specific adaptations to size and rotational speed as appropriate to the energy source or *prime mover*, but all are subject to the same set of analytical tools and well-understood behaviors. The basic idea is that any generator's *rotor*, or the part that rotates, is mechanically connected to something—typically by being mounted on a common rotating steel shaft—that continually exerts a torque and expends energy to make it turn.

The 21st century has seen rapidly growing deployment of renewable resources, particularly solar and wind, which introduced power electronics to the grid on a large scale. From the power systems perspective, this transition to inverter-based generation, more so than particular fuel choices, is a profound evolution.

The geography of resources is another important consideration. While fossil fuels and nuclear fuel can be transported quite easily to wherever it happens to be convenient and legal to build a power plant, most renewable resources require the infrastructure to come to them. For this reason, transmission lines are equally as relevant as the items in this chapter for solving the problem of a decarbonized electric grid.

15.1.1 Hydroelectricity

A hydroelectric turbine converts the kinetic energy from moving water into electrical energy in the rotation of the turbine-generator. This conversion process can be performed at high efficiency,

¹ An excellent text that provides much more detailed coverage, especially on solar, wind, and batteries, is Gilbert M. Masters and Kevin F. Hsu, *Renewable and Efficient Electric Power Systems* (Wiley-IEEE, 3rd edition, 2023).

since the mechanical energy in the water is already in a high-quality form, or state of low entropy, thermodynamically speaking. Water may be stored at an *intake* or *forebay* uphill from the plant and is fed toward the turbine through a *penstock* designed to provide a smooth and consistent flow. Over a river's downhill course, the same water may power multiple turbines.²

The power in moving water is related to its flow rate and velocity v , where m is mass and ρ (rho) is the density of water, 1000 kg/m³. The water's kinetic energy at the bottom of the penstock is equivalent to the gravitational potential energy P.E. associated with the elevation drop h and gravitational acceleration $g = 9.81$ m/s², minus friction losses in the pipe:

$$\begin{aligned} \text{P.E.} &= mg h \\ \text{K.E.} &= \text{P.E.} - (\text{friction losses}) = \frac{1}{2} m v^2 \\ \text{Power} &= \frac{\text{K.E.}}{\text{time}} = \frac{1}{2} \rho v^2 \cdot \frac{\text{Volume}}{\text{time}} \end{aligned} \quad (15.1)$$

Because friction losses are not obvious to calculate, a common convention is to account for them through an adjustment in the elevation difference h , known as hydrodynamic *head*, to reflect the *net head* (a smaller number). The power can then be calculated from the volumetric flow rate and the potential energy using the net head, without any need for information such as the penstock diameter, its surface roughness, etc. The gross head is the vertical difference between the water level in the forebay and the afterbay, or the elevation at which the water is released at atmospheric pressure.

Hydroelectric power plants have been built in a wide range of sizes, from kilowatt-scale run-of-river units to gigawatt-scale plants with dams and reservoirs for stored water supply (see also Section 15.3 on pumped hydroelectric storage). Different types of turbines are optimized for particular ranges of water pressure and flow, with the mechanical power conversion either drawing on just the water's velocity (*impulse turbine*) or also on the pressure difference (*reaction turbine*) with suction on the low-pressure side of the turbine blades.

Since the velocity of water is naturally limited, the generator has to compensate on the electromagnetic side to achieve synchronous speed (where 60 Hz equals 3600 rpm). This is accomplished in an all-analog manner with multiple magnetic poles (see Section 10.3.2).

From a system operations perspective, hydroelectric power plants with ample stored water behind them make an ideal renewable resource that is eminently controllable and fast-responding with well-understood characteristics. Hydro development is constrained, alas, by topography and water availability—with many of the world's best opportunities already taken—as well as the broader environmental impacts of creating reservoirs. Depending on the situation, these impacts can go far beyond aesthetics to include water chemistry, ecology, and population displacement.

Certain coastal sites may lend themselves to *tidal power*, a variation on the hydropower theme. Analogous to a hydroelectric reservoir, a tidal basin is constructed near the shore that is regularly refilled by the high tide. For this idea to make sense, the local tidal range must be very high. Another alternative is to submerge hydroelectric turbines on the seabed where there are strong tidal currents. Like for other renewable resources, the limiting factors in practice tend to be siting and permitting, rather than the availability of energy in principle.

15.1.2 Thermal Generation

The basic idea in steam generation is that water is turned to high-pressure steam inside some sort of boiler, and while being guided to expand through a set of fanlike turbine blades, the steam forces

² More than one hydroelectric development has dubbed its resource as “the hardest working water in the world.”

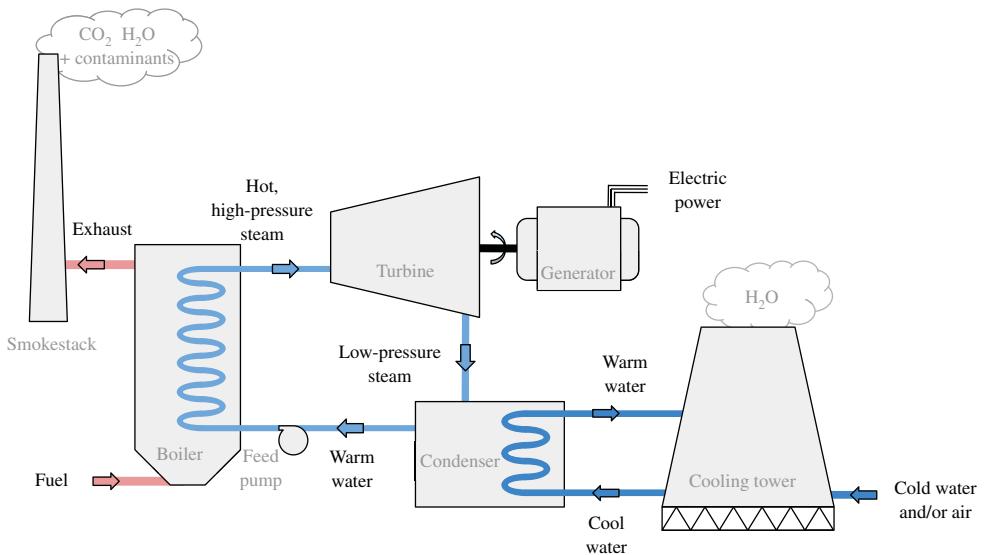


Figure 15.1 Conceptual illustration for a fossil-fuel power plant.

the turbine to rotate. Mounted on a shaft with an electric generator, the turbine spins the rotor (Chapter 10), and the rest is history. Before the advent of *fuel cells* that use an electrochemical process instead of combustion heat (Section 15.2), thermal generation was the only practical way to convert chemical fuel energy into electricity.

From the power systems standpoint, it is irrelevant what boils the water in a steam plant: it could be burning hydrocarbons, fissioning atoms, or concentrated sunshine. Figure 15.1 shows a simplified sketch of the steam generation cycle for a coal-fired plant. In reality, the plumbing of a power plant is much more complicated, often with several different stages of turbines optimized for successively lower pressure levels, and many different branches of pipe for recirculation and recapturing heat from the various stages so as to extract as much energy as possible. Here we stick to the conceptual basics.

After the hot, high-pressure steam pushes the turbine, it exits on the low-pressure side and is condensed back to liquid water in the *condenser*, creating suction on the back side of the turbine and thereby maintaining the state of disequilibrium that keeps the steam cycle in motion. The condenser is cooled by some outside source, which could be a steady flow of river or ocean water (*once-through cooling*), or *evaporative cooling* or just air convection with either *wet* or *dry cooling* towers. Once-through cooling uses high flow rates and a modest increase in water temperature of a few degrees (often limited by environmental regulations). Evaporative cooling uses much less water (owing to the high latent heat of vaporization³), but doesn't return it to the body of water from which it was extracted, so the environmental impacts depend on the circumstance. Dry cooling uses only air flow and recycles the condenser water, which is suitable for desert environments but less effective for heat transfer, and thus not as good for efficiency.

The crucial function of the cooling system and the condenser is not immediately intuitive—but consider that in its absence, the hot steam would have no reason to push from one side of the turbine to the other. Why is this important? The sheer volume of thermal energy transported as waste heat

³ The latent heat of vaporization for water is 540 cal/g at 100°C, while the specific heat of water is 1 cal/g·°C. Thus, to reject the same amount of heat, raising the temperature of cooling water by 5.4°C requires 100 times as much water as evaporating it.

comes into perspective when considering that most thermal power plants have efficiencies in the 30s of percent, meaning that two-thirds of the fuel energy ends up in the cooling water and only one third becomes electricity. This situation merits some discussion of the limiting factors.

When converting heat into mechanical energy or electricity, the second law of thermodynamics is paramount.⁴ Since thermal energy involves the random motion of many particles, it is inherently less ordered—that is, of higher *entropy*—than the mechanical kinetic energy of a macroscopic object, even though in both cases it is the sum of $\frac{1}{2}mv^2$ over all the particles. Because using thermal expansion to exert macroscopic force (say, on a turbine blade or piston) implies a redirection to increase order, it necessarily comes at an energetic loss. This can be visualized as the particles on the cold or low-pressure side pushing back and having irreversible work done on them, which ends up as waste heat. The conversion loss from thermal to mechanical energy is thus constrained by the comparative temperatures on the hot and cold side, regardless of the design details of the conversion device (where any such device is generically a *heat engine*), with the maximum theoretical limit of *any* heat engine given by the *Carnot efficiency* η_c (Greek lowercase eta):

$$\eta_c = 1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} \quad (15.2)$$

where it is very important to use *absolute temperature* (i.e., Kelvin).⁵ Equation (15.2) shows that any such process could be 100% efficient ($\eta_c = 1$) only if the cold side were at absolute zero, an impossibility. More practically, it shows that the efficiency of a steam turbine can be improved by either raising the steam temperature T_{hot} , or lowering the temperature T_{cold} of the medium where heat is rejected. In our case, this is the temperature on the low-pressure side of the turbine, which in turn depends on the temperature of the external cooling source.

On the high side, boiler temperatures are constrained by materials and cost. Reasonable boiler temperatures for coal-fired plants are in the neighborhood of 1000–1100°F. The upper end of the range applies to *supercritical* units, where liquid and gas phases of water coexist at very high pressure. These plants operate more efficiently but are more expensive to build.

On the low side, the limit is given by the ambient temperature of the environment or body of water. This explains why thermal power plants operate less efficiently in hot weather, and system operators have to take this de-rating into account when estimating available generation capacity on hot peak demand days.

A realistic power plant would be expected to perform at some fraction of the Carnot efficiency, considering that the turbine is not an *ideal* heat engine, and also that there are other conversion steps whose efficiencies might be in the 90s of percent (including the boiler transferring heat into the water and the generator converting mechanical into electrical energy). For example, using $T_{\text{hot}} = 1000^\circ\text{F} = 811\text{ K}$ and $T_{\text{cold}} = 100^\circ\text{F} = 311\text{ K}$ gives a theoretical Carnot efficiency of 61.7%, when a realistic plant at these temperatures might yield a little over half that.

⁴ The first law of thermodynamics says that energy cannot be created or destroyed in any conversion process, or “You can’t win.” The second law says that any spontaneous process is associated with increasing entropy, a measure of disorder in a collection of physical objects or particles. Order can be increased (and entropy decreased) locally, but there is an energetic price to pay. In other words, “You can’t break even.” More second-law aphorisms include “you can’t unscramble an egg,” and “the papers on your desk will not alphabetize themselves.” The third law says that entropy is zero only at absolute zero temperature (0 K), where nothing moves and humans don’t exist. Thus, “You can’t get out of the game.”

⁵ Kelvins have the same degree increment as the Celsius scale, offset by 273.15. Thus, $0^\circ\text{C} = 273.15\text{ K}$ and absolute zero is $0\text{ K} = -273.15^\circ\text{C}$. Taking a meaningful ratio of temperatures requires using an absolute scale that is zero at absolute zero. Note that Kelvins are an SI unit, and it is technically improper to say “degrees Kelvin.”

While the Carnot efficiency imposes hard constraints on turbine efficiency, the overall systemic efficiency of fuel utilization can be improved by combining stages that are each optimized differently. *Combined cycle* plants use both steam and *gas turbines*, which resemble jet engines in that the turbine blades are propelled directly by the combustion exhaust gases. Small gas turbines are sometimes used for peak power generation, since they can ramp power up and down much more quickly than a steam plant. Gas turbines by themselves are less efficient and thus costlier to operate, but when recovering the exhaust heat for a steam turbine stage, the combined efficiency is greater.

Even better systemic gains can be made by directing the inevitable waste heat toward some useful purpose, in a *combined heat and power* (CHP) facility, also called *cogeneration*. For example, the rejected steam from a turbine may still be plenty hot enough for some industrial process, and certainly for water and space heating in commercial buildings, or even *district heating* for the residential sector; the crux is the steam delivery infrastructure. In these cases, the electric generation is called a *topping cycle*. Alternatively, for industrial processes that require extremely hot steam, cogeneration can be added at the low temperature end—with less turbine efficiency, of course, but better than wasting the steam entirely—as a *bottoming cycle*.

In general, a heat engine may operate with a different working fluid than water (e.g., ammonia or other refrigerants), over different temperature and pressure ranges. The basic design requirement is that the working fluid be evaporated by the heat source, and condensed in the location where waste heat is to be rejected.

An instructive case is *ocean thermal energy conversion* (OTEC), where the modest but consistent temperature difference between warm surface water and cold deep water drives a heat engine cycle to generate electricity. This technology has been demonstrated on tropical islands. The Carnot efficiency for OTEC is just a few percent at best, requiring very high flow rates of water (with large piping and heavy pumping loads), but the energy reservoir is practically unlimited.

15.1.2.1 Fossil Fuels

Historically, the overwhelming majority of thermal generation has been fueled by coal, followed by natural gas in the later part of the 20th century; petroleum fuel oil has also been used. All fossil fuels are essentially hydrocarbons (of the general formula $C_x H_y$) whose intrinsic chemical potential energy lies in the absence of oxygen; they release energy when forming carbon dioxide and water. Because of the greater ratio of hydrogen to carbon in methane (CH_4) as compared to coal (almost all carbon), the combustion products of natural gas (mostly methane) include less CO_2 and more H_2O per unit of chemical energy, a favorable balance in terms of greenhouse emissions.

Coal burning in particular is also associated with greater toxic atmospheric emissions due to contaminants, especially sulfur oxides (SO_x) which can be *scrubbed* from smokestacks at some expense; nitrogen oxides (NO_x), which are a side effect of any fuel combustion (due to nitrogen in the air) depending on the temperature of the flame; and particulates or unburned hydrocarbons of many varieties. While natural gas is often considered a “clean” fuel, its combustion is still imperfect and it is not carbon neutral unless it comes entirely from above-ground sources (such as biogas fermentation or other waste), or all its carbon is captured pre- or postcombustion.⁶

15.1.2.2 Biomass

A broad range of biofuels in solid, liquid or gaseous form can be used for steam generation. The economics of biomass power plants tend to favor smaller facilities that leverage local niche

⁶ Note also that methane itself is a more potent greenhouse gas than CO_2 , so that any leaks in the natural gas infrastructure are consequential for climate protection.

opportunities (e.g., burning leftover wood chips or walnut shells). Constraints on economic scalability of biomass electric generation include fuel collection effort (for which there will be some maximum practical radius), and competing higher-value uses for cropland. In the larger context of decarbonization strategies, biofuels may have higher value for uses where the ease of fuel storage and transport has important advantages over electricity as an energy carrier.

15.1.2.3 Geothermal Power

Another possible source of steam is *geothermal*.⁷ The highest quality geothermal resource is *dry steam*, which can be fed into a steam turbine almost directly as it comes out of the ground, with some chemical cleaning (e.g., removing sulfur) both to meet environmental regulations and to prevent corrosion in the turbine. More commonly, geothermal reservoirs deliver hot water, which then requires additional heat exchange steps for steam generation. Although there is some heat underground everywhere if one drills deep enough, the practicality of geothermal power generation depends on easy access to high temperature fluids, and also on what toxic chemicals accompany these fluids at a particular site. Geothermal power plants have operated successfully for many decades, but in limited choice locations. *Fracking* technology developed for oil and gas extraction may contribute to expand the range of future geothermal applications.

15.1.2.4 Nuclear Power

From a power systems perspective, nuclear fission is just another way to boil water. The turbine-generator components of a nuclear facility are very similar to those of conventional fossil-fuel plants, although the plumbing behind them is more complex and engineered to different safety standards. The steam temperatures and pressures used in nuclear plants are characteristically lower than in fossil-fuel plants, implying some changes to turbine design as well as a lower operating efficiency. Of course, energy conversion efficiency is not a salient point when fuel energy is plentiful and its marginal cost represents a small fraction of the total facility cost.

The source of nuclear fission energy is the conversion of mass into thermal energy according to $E = mc^2$ when a *fissile* isotope such as uranium-235 breaks into pieces including *fission products* and several individual neutrons.⁸ When fission neutrons encounter another fissile nucleus at the appropriate speed, they can provoke another fission in what becomes an exponentially propagating chain reaction. In a reactor (unlike a nuclear bomb), the reaction rate or *criticality* is precisely managed by controlling the amounts and spatial distribution of fissile nuclei in the fuel, *moderators* that collide with neutrons to slow them down and thereby enhance the reactivity, and neutron *absorbers* (e.g., in the *control rods*) that stop the reaction.

The most widely used reactor fuel is *enriched uranium*, where enrichment means increasing the concentration of the U-235 isotope from its natural occurrence (0.7%) among the more abundant

⁷ Depending on one's definition of "renewable," it may or may not apply to geothermal resources. On one hand, the heat below the earth's crust is an unfathomably large and for all practical purposes inexhaustible reservoir; it also is continually replenished by natural radioactive decay. On the other hand, a specific geothermal resource involves transport of heat to the surface by some flow of water, natural or pumped. If the rate at which heat is removed from the hot subsurface rocks exceeds the rate of replenishment, a local geothermal well or reservoir can be depleted.

⁸ While the number of nucleons (protons and neutrons) is conserved in the fission process, the sum of their masses is ever so slightly less in the product configuration than in the original bulky nucleus, owing to the *binding energy*. This type of discrepancy is never encountered in chemical processes, where everything revolves around electrons while the nuclei maintain their elemental identity, so that mass is strictly conserved without having to call it mass-energy. Because the speed of light c is a large number, a tiny amount of mass m corresponds to a large amount of energy E . Fission products tend to be highly radioactive due to their imbalance of protons and neutrons, since the stable ratios differ for atoms of different size. This radioactivity also contributes significantly to a reactor's heat production.

U-238, typically to around 4%. A majority of countries with civilian nuclear programs, including the United States, use a *once-through* fuel cycle.⁹ The most widely used technology for civilian energy production globally is the *light water reactor* (LWR) technology, which includes two types: *boiling water reactors* (BWRs) and *pressurized water reactors* (PWRs) that were originally developed to power Navy ships and submarines.¹⁰ The water in an LWR serves both as a moderator and as a coolant to transfer heat to the steam turbine cycle. In a PWR, the water circulated through the reactor core is separated from the steam cycle by a heat exchanger called a *steam generator*, whereas the water/steam mixture in a BWR circulates directly through the core. Other reactor designs that are not in commercial operation at the time of this writing use different coolants (such as gas, molten salt, or liquid metal).

LWRs are periodically shut down for *refueling outages* (say, every 18 or 24 months), usually independent of the energy production history. Under typical utilization or *burnup*, the energy density of uranium is four to five orders of magnitude greater than for chemical fuels, with one fuel pellet (about 1 cm³) equivalent to about a ton of coal. Though comparatively small in volume, spent reactor fuel and the question of how to safely and ethically dispose of it has been controversial in the United States and other countries.¹¹

Operational constraints for nuclear power plants center on the imperative to prevent the release of radioactive materials to the environment, which can result from a loss of coolant or melting of the reactor core. Due to the radioactivity of fission products, the core and any spent fuel continues to produce heat long after the fission reaction has stopped and must be actively cooled for years, at an exponentially declining rate, to prevent melting and consequent dispersal.

It is technically possible to make rapid and precise changes to the fission reaction rate and thus the reactor power output by moving control rods, adjusting the water chemistry, or adjusting the pressure; indeed, *load-following* operation is essential on Navy ships, and common where a large fraction of generation capacity is nuclear (e.g., in France). However, reactor operators generally eschew unnecessary changes in power level. Reasons include avoiding mechanical and thermal stress fatigue due to expansion and contraction, avoiding delicate issues around maintaining stable reactivity,¹² and minimizing opportunities for any disturbances or mistakes. If responding to grid needs, nuclear operators may conservatively constrain the power ramp rate (in MW/min), or limit the range of power adjustments to a few percent of rated output. Given that the marginal cost of energy produced is negligible once the reactor is running, there is also a strong economic incentive to operate at full rated capacity. U.S. civilian nuclear plants have been categorically operated as baseload units, with steady output as much as possible.

In the event of a grid disturbance that trips the generator, a rapid reactor shutdown or *scram* occurs, followed by cooling with off-site power, or on-site emergency generators as a last resort. To minimize the chances of a loss of off-site power, U.S. regulations require two separate transmission

9 By contrast, a *closed fuel cycle* means spent reactor fuel is *reprocessed* to extract remaining fissile isotopes such as plutonium-239, which is a byproduct *bred* by exposing U-238 to the reactor milieu. This technology is central to the production of nuclear weapons.

10 The term *light water* distinguishes regular from *heavy water* that contains the deuterium isotope of hydrogen, used in some other reactor types (e.g., with unenriched uranium) to reduce neutron absorption by hydrogen.

11 While most fission products decay quickly, spent fuel also contains hazardous long-lived isotopes that impose unique design criteria for disposal sites.

12 The main issue for LWRs is the accumulation of xenon-135, a fission product that strongly absorbs neutrons and thus reduces reactivity, but decays with a half-life of about nine hours. Xe-135 concentrations undergo transients for several hours after a significant change in power, requiring compensation with other controls; this motivates operating rules to maintain steady output for a minimum time period after load following maneuvers. The effects of Xe-135 accumulation become more pronounced as uranium fuel approaches the end of its refueling cycle, setting a limit for minimum stable load without losing criticality.

links to each nuclear plant. Nuclear reactors are unique among generation resources in that they have their own stake in a stable and reliable electric grid—both to guarantee receptivity for their electrical output, and to provide for active cooling—not just for economic reasons, but for safety.

A high priority for innovative reactor design is increased operational flexibility, along with passive safety (i.e., not requiring actively powered cooling, but relying on natural convection to keep the fuel below its melting point). As of the early 2020s, there is investment in new technologies including smaller, modular reactors (ranging from ca. 10–300 MW as compared to the typical 1000 MW for LWRs), but there is not yet an established commercial track record.

Nuclear fusion, a very different physical process than fission, is not at a stage of development relevant for this book.

15.1.2.5 Concentrating Solar Power

Solar thermal power plants, also known as *concentrating solar power* (CSP), use mirrors that focus sunlight onto a receiver containing a heat transfer fluid, which then produces steam through a heat exchanger to power a fairly conventional turbine generator. The sunshine is yet another way to boil water. This technology and operating principle is very much distinct from photovoltaics (PV) (Section 15.1.3).

Several different configurations have been built, and some operated commercially for decades: central tower, parabolic trough, and parabolic dish. The common idea is that the mirrored surface, which may consist of many separate pieces, approximates a parabolic shape with its focal point at the receiver.¹³ In the case of the central tower, large sun-tracking mirrors or *heliostats* are mounted on the ground surrounding the tower. This design approach banks on economies of scale (100 MW and above), and taking advantage of built-in thermal storage to provide a somewhat dispatchable generation resource. The parabolic trough design as commercialized in the 1980s uses a linear focus with a translucent pipe that carries the heat transfer fluid, aggregating to medium-size generation units (tens of MW) also with a thermal storage option. This geometry requires only single-axis tracking. The parabolic dish is a smaller (several kW) version that resembles a satellite dish, where the whole unit tracks the sun and some type of heat engine may be mounted right at the receiver.

A key metric for CSP is the *concentration ratio*, expressed in units of *suns*, that is the ratio of the mirrored collector area to the receiver area onto which that sunlight is focused. Concentration ratios range from the 100s to over 1000. Considering that one full sun irradiance is 1000 W/m^2 , the engineering for materials and safety is far from trivial (and flying through the concentrated beam is certainly fatal for birds). A higher concentration ratio generally means a higher operating temperature, which is advantageous for thermodynamic efficiency, although it also implies greater losses if the heat transfer fluid is transported some distance. Conversion efficiencies from sunshine to electricity are reported up to 35%, with 25% being more typical. The technology continues to evolve, although it is challenged to compete with low-cost PV. Focusing sunlight also means that only the direct, not the diffuse component can be utilized, which strongly favors climates with low humidity (where, in turn, cooling water is more likely to be scarce).

Commercial solar thermal plants have used natural gas as a supplementary energy source for increasing the plant capacity factor, improving unit availability in variable weather, and for morning startup. While advantageous for economics and reliability, this option is controversial in terms of disqualifying CSP from being a purely clean and renewable resource. Despite these issues, it is important to note that solar thermal power exists as a proven technology option that is fully compatible with conventional steam generation.

¹³ Mathematically, a parabola can be defined as a curve that will reflect incoming parallel rays to a single point, and *vice versa*; this is why parabolic dishes are used for communications.

15.1.3 Solar Photovoltaics

PV is what most people mean by the generic term “solar power.” Developed in the mid-20th century to power satellites, and subsequently adopted for growing numbers of niche and remote applications, this technology has become ubiquitous as its costs declined dramatically. Grid-connected PV installations for homes and utilities emerged in the 1980s. Scalable over many orders of magnitude, with no moving parts and no particular hazards or vulnerabilities, PVs are eminently robust and adaptable to an enormous variety of circumstances. There is no clear-cut constraint on suitable climates; simply, the worse the solar resource, the more PV and the more area it will take to meet a given load. Efficiencies of commercial PV modules are typically in the range of 10–20% as of this writing. Higher efficiencies are possible but not usually cost-effective for practical installations, other than high-value uses with limited available area.

PV cells consist of a specially treated semiconductor material, most often silicon, that produces an electric potential when exposed to light (the *photovoltaic effect*). Incoming bits of light (photons) interact with electrons inside the semiconductor, making them free to travel. The special treatment of the PV material creates an asymmetry called a p-n junction which compels free electrons to move in one direction, toward the already negative terminal (see Section 14.3.1). This asymmetry is never used up; the work done by the moving electrons comes entirely from the light energy. Many different PV technologies exist, both on the market and under development, using different materials and manufacturing processes.¹⁴

The cell voltage is determined by the *bandgap* of the semiconductor material, which represents the amount of energy per electron that is required to elevate it into the conducting state (called the *conduction band*), and that conversely becomes available (minus losses) when the electron returns to its bound state, doing work on the external circuit along the way. Any photon with a minimum energy corresponding to the bandgap is capable of creating an electron–hole pair.¹⁵ Photons with insufficient energy (too long a wavelength) pass right through the material or just heat it up; this is called *red loss*. Photons with more than enough energy still only create one electron–hole pair apiece, and the extra energy is dissipated in the material without doing useful electrical work; this is called *blue loss*. Clearly, there is a trade-off that limits the maximum possible conversion efficiency: high-bandgap materials will have more red loss, and low-bandgap materials more blue loss. A clever way around this dilemma is to combine materials of different bandgaps into *multi-junction cells*, with the high-bandgap material on top.

A typical open-circuit voltage (i.e., under no load and thus with minimal losses) for a plain individual silicon cell is about 0.6 V. The open-circuit voltage decreases slightly with temperature (essentially, due to entropy within the cell), so that PV systems generate the most power on clear and sunny but cold days. The short-circuit current, provided by the cell when there is no impedance, is directly proportional to the amount of sunshine, or the number of photons striking the cell.

PV cells are aggregated into panels or *modules* of practical size and useful voltages, where stringing cells together in series increases the voltage and adding cell area or strings in parallel increases

¹⁴ Some of these processes, like semiconductor manufacturing in general, involve toxic substances. Since PV cells have lifetimes of decades, there is no data yet on disposal or recycling issues at high volume. There is no shortage of silicon, and modern processes have significantly reduced the materials volume and manufacturing energy required so that the *energy payback* is easily within a year.

¹⁵ See Section 1.5.6 on electromagnetic radiation. The energy E carried by a single photon is given by $E = hf$, where h is Planck's constant (a very small number, $h = 6.626 \times 10^{-34} \text{ m}^2 \cdot \text{kg}/\text{s}$) and f is the frequency, which is related to the wavelength λ by the speed of light $c = \lambda f$. Visible light ranges from about 380 nm (violet) to 700 nm (red). The bandgap of silicon is 1.1 electron volts ($1 \text{ eV} = 1.6 \times 10^{-19} \text{ J}$), which corresponds to infrared light of $\lambda \approx 1100 \text{ nm}$. Therefore, all of the visible spectrum works for silicon PV cells, but much of the energy from sunshine comes from the infrared part of the spectrum.

the available current. Common d.c. open-circuit voltages are around 15–20 V, which is sufficient to charge a 12-V battery. Common module sizes are on the order of several hundred watts.

The *operating characteristic* or *IV curve* of a PV cell or module, discussed in Section 2.5.1 and illustrated in Figure 2.9, shows the possible combinations of voltage and current for a given set of conditions, most importantly depending on the *insolation* or *solar irradiance* in W/m². The actual operating point on the IV curve is determined by its intersection with whatever load is connected to the terminals.¹⁶ The *maximum power point* (MPP) corresponds to the maximum possible instantaneous product of voltage and current (which is less than $I_{sc} \cdot V_{oc}$ due to the shape of the curve). Modern inverters use MPP tracking to continually vary the operating voltage (by effectively varying the load impedance as seen by the PV module) in such a way that always extracts the most power available, even as insolation changes.

An installation with multiple PV modules is called an *array*, which can range from the size of a rooftop to covering many acres. Modules are connected in series with additive voltage to form strings, and strings are connected in parallel with additive current. When a *string inverter* is used, the voltage of each string of modules must be matched to the inverter's d.c. input parameters, under all possible temperatures. Common sources of systemic loss include partial shading and voltage mismatch, as a reduced output from any one module will compromise the ability of the entire system to operate at its MPP. Alternative designs use *microinverters* or other custom voltage conversion devices attached to each module, to make systems more tolerant of partial shading or failure of individual modules.

When designing or using any solar energy installation, it is useful to be aware of *solar geometry* that describes the position of the sun in the sky as a function of latitude, time of day, and time of year.¹⁷ Although nonconcentrating PV modules can use diffuse light from all directions, the incident sunshine should ideally be close to *normal* (i.e., at right angles) to the plane of the solar collector. When incident light is off-normal by an angle θ , the power intercepted per unit area is reduced by a factor of cosine θ . Additional loss will be due to the surface becoming more reflective at shallower angles of incidence.

In choosing the orientation of fixed arrays, some compromises are clearly necessary: one might maximize annual energy production, seasonal production, or revenue, if not simply following the orientation of an available roof. The tilt angle will typically be similar to latitude, with a steeper tilt favoring winter generation. The preferred azimuthal orientation may be somewhat westerly, since the value of electricity to the grid is typically greater in the late afternoon when demand tends to peak. Thus, a west-facing solar plant might successfully take advantage of high afternoon rates, at the expense of total energy production. This option helps to address the *duck curve* discussed in Section 6.4.

¹⁶ For example, a simple resistor has an operating characteristic determined by Ohm's law, which is a straight line through the origin described by $V = IR$, where the resistance R determines the slope.

¹⁷ Solar geometry is delightfully deterministic and exact. While modern smartphone apps can chart the path of the sun across the sky for any given place, date, and time zone, some readers will feel satisfied to know directly the two basic equations for the sun's altitude and azimuth angles, as a function of latitude L , hour angle H , and *declination* δ . L ranges from 0 at the equator to +90° at the North and -90° at the South Pole. The time of day is defined by convention as 0 at solar noon (when a shadow is shortest), positive in the morning and negative in the afternoon, with one hour equivalent to 15°; for example, $H = +45^\circ$ means 9 a.m. solar time. The conversion from clock time to solar time is a separate process that involves longitude relative to the reference meridian, the *Equation of Time* correction for the tilt of the earth's axis and the eccentricity of its orbit (with a range of roughly ±15 minutes), and daylight saving time if applicable. The declination, which is always the same everywhere on the planet, ranges from +23.45° in June to -23.45° in December and is zero at the equinoxes. It is given by $\delta = -23.45^\circ \cos[360(d + 10)/365]$, where d is the day of the year. With these definitions, the sun's altitude β is given by $\sin \beta = \cos L \cos \delta \cos H + \sin L \sin \delta$, and its azimuth ϕ (zero due South and positive to the East) is given by $\sin \phi = \cos \delta \cos H / \cos \beta$.

15.1.4 Wind Power

Wind energy has been used in mechanical applications (sailing, milling, water pumping) for millennia; the first conversion of wind power to electricity dates back to the 1880s. Wind turbine projects up to the megawatt scale were built in the early and mid-20th century (with Denmark as a global center of expertise), but rapidly expanding fossil fuel use eclipsed commercial interest in the resource. Wind generators served mostly niche applications until a renaissance of grid-connected wind development in the 1980s, with initially slow adoption leading up to explosive growth worldwide around the 2010s.

The technological development of wind power over the course of the century could be characterized as incremental improvements, since the overall shape and function of the most common turbines—with horizontal-axis, three-blade rotors—has stayed remarkably consistent.¹⁸ However, big changes have been made in materials and manufacturing to produce extremely strong and durable mechanical components at low cost, as well as in electronic power conversion and digital control. A third area of ongoing development is offshore wind turbine installations for deep water.

The mechanical power in the wind is completely analogous to that in flowing water, per Eq. (15.1), with the key difference that the density ρ of air is three orders of magnitude less than water.¹⁹ The volumetric flow rate of air is the product of the *swept area* A of the wind rotor and the wind speed v , which gives the rate at which air traverses that area to interact with the rotor blades. Since this speed is the same v as in the kinetic energy, the power in the wind is given by the simple formula²⁰

$$P_{\text{wind}} = \frac{1}{2} \rho v^2 \cdot A v = \frac{1}{2} \rho A v^3 \quad (15.3)$$

Since the air cannot be brought to a complete standstill by the rotor (lest it block the arrival of more air, which would simply go around the rotor), there is an upper limit to the fraction of this power that can be extracted, called the *Betz limit*; this theoretical limit is 16/27 or 59.3%.²¹ Commercial wind turbines today typically get in the neighborhood of 40% and, in good conditions, up to 50% mechanical conversion efficiency.

Wind turbine blades have an airfoil shape analogous to sails and airplane wings, designed to be pulled by *aerodynamic lift* rather than pushed by *drag* forces. As counterintuitively as one can sail upwind, the rotor blade speed through the air may exceed the wind speed relative to the ground and rotor plane; their ratio at the tip of the blade is called the *tip speed ratio*. Adjusting the *pitch* of the rotor blade changes the *angle of attack* of the *apparent wind* relative to the blade. Just as in an airplane, a steeper angle of attack increases the lift force—up to the point of *aerodynamic stall*, where the lift quite suddenly succumbs to turbulent air flow. Modern wind turbines can use *pitch control* to continually optimize lift, spill excess power in high winds, or deliberately induce stalling.

Wind speeds are notoriously variable. Since the power in the wind is a function of velocity cubed, wind power varies over a wide range, and very dramatically at higher wind speeds. In the design of wind turbines, there is an inevitable trade-off between high efficiency at low wind speeds, and the

¹⁸ Numerous other designs, including vertical axis turbines, have been built and operated commercially and each has distinct advantages, but none has proven as universally suitable and successful under diverse conditions as the three-blade horizontal axis rotor.

¹⁹ The density of air is $\rho = 1.225 \text{ kg/m}^3$ at 0°C and 1 atmosphere pressure.

²⁰ Equation (15.3) is also known as the *fan law* when referring to the power requirement for blowing air—which means imparting kinetic energy to it. The fan law explains why variable speed motor drives in ventilation equipment can save such a surprising amount of energy: because of the cubic dependence on speed.

²¹ The Betz limit can be derived with some calculus, maximizing the difference in kinetic energy contained by a volume of air before and after passing through the rotor. It does not apply to a hydroelectric turbine in an enclosed space, where the water has nowhere else to go and power can be extracted from the pressure difference, while the wind rotor is in an open space.

ability to safely extract power at high wind speeds. The key is matching the particular machine to the statistical wind speed distribution at a given site.²² Any given machine has a *cut-in speed*, below which it is not worth operating the generator, and a *cut-out speed*, above which it is unsafe and the brakes are applied. The full rated output of the generator is achieved at the *rated wind speed*. At wind speeds between rated and cut-out speed, some power must be spilled (e.g., by blade pitch control).

The crucial advantage of the offshore wind resource is that wind speeds on the sea are not only much higher on average, but they are also more consistent, allowing for more targeted optimization. Offshore wind power means engineering very robust turbines and towers that are either solidly anchored on the sea floor, or installed on floating platforms in deep water. Undersea d.c. electrical cables are standard.

The early 2000s have seen a trend toward larger wind turbines, especially offshore, with rotor diameters above 100 m (bigger than a football field!) and generator ratings of several MW for each machine. Greater *hub height* takes advantage of increasing wind speeds with elevation (known as *shear*), and greater swept area means more power for fewer towers installed. One challenge for large horizontal-axis rotors is that the controls must compensate for the difference in wind speed at the top and the bottom of the rotation. A simple practical limitation is the transport of very long and rigid turbine blades to location, on a truck with finite length and turning radius.

To maximize mechanical power conversion efficiency over a range of wind speeds, it is necessary to maintain a constant angle of attack—either by varying blade pitch, varying rotational speed so as to maintain a near constant tip speed ratio, or some combination. Therefore, a key aspect of the electrical design for wind generators is to allow variation in rotational speed, while producing alternating current of a constant frequency.

Early grid-connected wind generators relied primarily on simple and inexpensive induction generators; these are now referred to as Type I. Induction machines have a few percent of speed variation over their range of torque and power output known as *slip* (see Section 10.6), but this serves more as shock absorption than optimization for wind speed changes.

Induction generators have two important limitations: they are not capable of controlling bus voltage or reactive power (VAR) output. Rather, they always “consume” VARs in the course of injecting watts to the grid. Reactive power can be compensated with capacitors (see Section 3.4.5), at a cost. Also, induction machines are not capable of controlling a.c. frequency, or starting up without an a.c. signal already present at the interconnection point. Although this poses no particular problem as long as the *system penetration* or percentage contribution to the grid is small, it means that induction generators cannot be considered a fully controllable resource on par with synchronous machines, and they cannot be used for system restoration or *black starts*.

Induction machines can be designed for a wider range of speed variation with respect to power; these are known as *high-slip* or Type II generators. Another variant is the *doubly fed induction generator* (DFIG), or Type III machine. In the DFIG, power electronics are used to route some output power back to the rotor and adjust the electrical frequency of the rotor current.²³ This provides a wider range of mechanical rotation speed while maintaining a constant a.c. frequency in the stator

²² Absent detailed measurements, a standard assumption is that the wind speed statistics follow a *Rayleigh distribution*, which is uniquely characterized by an average speed. But since microclimates can vary greatly, empirical local wind data collection is invaluable.

²³ Recall from Section 10.6 that the stator (output) a.c. frequency is always the sum of the rotor electrical frequency and its mechanical rotation rate. In a synchronous machine, the rotor current is d.c. By introducing a variable frequency of the rotor current, the DFIG allows changing the rate of mechanical rotation while maintaining a constant frequency a.c. output.

windings; it also allows power factor control. The basic design idea is that since power electronics are expensive, it is advantageous to route only a fraction of the machine's overall output power through them.

With decreasing cost of power electronic conversion, Type IV wind generators have become standard. These initially produce *wild a.c.* at whatever frequency is mechanically suitable, rectify this output to d.c., and then generate well-behaved a.c. with an inverter. This affords rather complete control over reactive power, voltage, and waveform. The electrical conversion losses are dwarfed by the mechanical efficiency gains.

15.2 Distributed Generation

Distributed generation (DG) describes electric power generation that is geographically distributed or spread out across the grid, generally smaller in scale than traditional power plants and located closer to the load, and often on customers' property. As an increasingly prevalent set of technologies, DG is associated with interesting questions about the overall design of the grid, operating strategies, economics, environmental impact of electricity production, and, ultimately, energy politics. This section is intended as a brief and general introduction to the main characteristics of these technologies and their integration issues, from a power systems perspective.

15.2.1 DG Resources

Important DG technologies today include solar PV, fuel cells, microturbines, and occasionally wind turbines. By definition, these differ from traditional energy resources in the scale and range of suitable locations for their deployment. PV in particular epitomize the idea of a *modular* technology, in that a PV system can be built at just about any scale, from pocket calculator to a megawatt array. Large arrays are made by simply combining more modules in parallel to obtain arbitrarily large amounts of current and power.

Economies of scale for PV have mostly to do with volume purchasing, fixed design and construction costs, and other infrastructure-related costs; they are not intrinsic to the technology. While it is true that a large, centralized solar farm can usually produce electricity at a lower cost than the sum of many small installations, the economies of scale are much less pronounced than for steam generation. In fact, there may be diseconomies of scale associated with siting and licensing, or the need to upgrade transmission infrastructure. Thus, PV is suitable for smaller, location- and load-specific applications, where the localized benefits often far outweigh the somewhat higher cost of energy.

From the perspective of the electric grid, important characteristics of DG resources include the electrical properties of the generator component, and the timing of electricity production. PV and fuel cells naturally produce direct current, and accordingly interface with the grid through inverters (Section 14.4).

15.2.1.1 Fuel Cells

Fuel cells resemble flow batteries in that a chemical reaction forces electrons to one side, where the reactants, hydrogen (H_2) and oxygen (O_2) gas, are continually supplied from an external source. The electrical energy is released from the chemical reaction where hydrogen and oxygen combine to form water, which is safely released into the atmosphere. Although hydrogen as an energy storage medium can be produced by the reverse process—electrolysis—from water and any electrical energy source, most commercial hydrogen today is obtained from natural gas in a process called

reforming. Some types of fuel cells incorporate the reforming step and use methane directly as a feedstock. In that case, there is also carbon within the waste products, which may be released as CO₂ or captured in some form.

The electrochemical process for extracting the fuel energy is more efficient in principle than burning the gas to drive a turbine because it avoids the fundamental thermodynamic limitations of converting heat into motion. Thus, while combustion engine efficiencies are more commonly in the range of 30s of percent, fuel cells can attain in the neighborhood of 60% conversion efficiency from chemical to electric energy. A variety of fuel cell technologies using different electrolytes and catalysts are available and at different stages of commercial maturity.²⁴ In practical terms, the technologies differ in their operating temperature and size ranges, from kW to MW. Thus, some types are suitable only for installations such as commercial buildings rather than, say, residential use, but they are still much smaller than a traditional power plant.

Like for PV cells, the d.c. voltage supplied by a fuel cell is determined by the properties of the materials and is on the order of 1 V per cell. Multiple individual cells are connected together in series to form *fuel cell stacks* with a convenient operating voltage to supply the inverter, whose parameters and control settings are the main observable characteristics from the grid perspective.

15.2.1.2 Microturbines

Microturbines are powered by natural gas, essentially methane (CH₄), which may be derived from renewable sources such as wastewater, landfill, or manure digester gas. They are considerably smaller than steam turbines or even gas turbines, with units in the range of tens to a few hundred kilowatts that can readily fit into a basement. Although atmospheric emissions from microturbines tend to be higher, they can provide systemic efficiency gains by way of a *cogeneration* or *combined heat and power* (CHP) option. This allows making simultaneous use of electricity and waste heat, on a smaller scale than conventional cogeneration facilities that use steam in industrial settings. Microturbines tend to operate at very high rotational speeds, which means that their a.c. output must be adapted to the grid frequency.

15.2.1.3 Small Generators

There is of course a broad variety of more basic generators that use diesel, natural gas or propane as fuel, ranging in size from single-digit kW to MW. However, most of these are used exclusively for backup generation: they are not operated on a routine basis, and they never inject power into the grid. Rather, they are *grid-forming* devices with synchronous generators designed to maintain a specific, isolated power island during emergencies. Installed for the purpose of providing dependable service to critical loads, the economic cost–benefit calculation for backup generators is of an entirely different nature than for grid-connected resources, and operating efficiency as well as emissions will be less of a concern.²⁵ Even if backup generators could compete with retail electricity on a per-kWh basis, environmental regulations may preclude their operation for the purpose of feeding into the grid. One interesting exception involves mobile diesel generators (e.g., on a trailer) used by a utility to temporarily energize a local distribution circuit during maintenance or restoration work.

²⁴ Technologies include polymer electrolyte membrane (PEM), alkaline fuel cell (AFC), phosphoric acid fuel cell (PAFC), molten carbonate fuel cell (MCFC), and solid oxide fuel cell (SOFC). See <https://www.energy.gov/eere/fuelcells/articles/fuel-cells-fact-sheet> (accessed March 2023).

²⁵ See Section 13.2.2 for a brief discussion of valuing service reliability. Noteworthy emissions from fossil fuel combustion include carbon monoxide (CO), nitrogen oxides (NO_x), unburned hydrocarbons, and particulates, especially for diesel.

15.2.1.4 Small Wind Turbines

Wind turbines are among the oldest small-scale electric generation technologies. Kilowatt-scale wind machines don't offer the efficiency and competitive energy cost of large, state-of-the art turbines, but they are an entirely proven and viable technology. The main practical constraints are local wind resource and permitting. It is generally much easier to find a sunny than a consistently windy spot, and variation of wind resource with substantial uncertainty on seasonal, daily and shorter time scales means less utilization of installed capacity. The intrinsic hazard of moving rotor blades (which can conceivably break or detach) makes wind turbines unsuitable for densely populated areas.

15.2.2 DG Integration

Beyond being technically viable and affordable on a small scale, a key requirement of DG to be sited very close to end uses is that it must be environmentally compatible. The easiest technology to site is PV, which brings minimal nuisance: zero emissions, no noise, minimal aesthetic impact, and options to integrate power installations with buildings or other structures (e.g., shaded parking lots). Without any moving parts whatsoever, the maintenance requirements for PV systems are also minimal, allowing the technology to be installed in remote locations. Fuel cells require more supervision, but they are not too hazardous to operate in occupied buildings. The higher power density of fuel cells (owing to the fuel's energy content) makes for compact units compared to solar generation, and installations in spaces such as office basements are becoming standard. Siting considerations for microturbines are similar to fuel cells, except that the emissions from combustion include more hazardous components.

Siting generation close to loads rather than in a centralized manner has significant technical implications for the grid. Many of these impacts are positive at least in theory, but they also introduce complications when interfacing with the legacy infrastructure. Predictability and controllability are big factors, and the effects can be very different depending on the amount of DG present, or *penetration level*.

First, we should expect DG to generally reduce I^2R line losses throughout the grid (see Section 3.2). Although a quantitative estimate of loss reduction due to DG requires an explicit power flow analysis comparing specific scenarios, we can say qualitatively that when generation occurs next to a load of comparable size, effectively negating this load, it would tend to have the effect of lowering current flow in the transmission and distribution lines that connect this load to major generation sources in the grid. Thermal energy losses are then reduced in proportion to the square of the current and the resistance of all the affected lines.

Second, DG can offer VAR support, potentially offsetting the need for other devices such as capacitors and voltage regulators (see Section 7.4). Note that while much of the grid's overall reactive power demand is met inexpensively through centrally located synchronous generators, some portion of it must be injected locally, that is, near loads, in order to maintain an acceptable voltage profile throughout the distribution system. If DG can provide voltage support in a reliable manner such that other hardware installations or upgrades can be avoided—for example, it might obviate the need for a new capacitor bank, or it might reduce wear on a load tap changer—there are systemwide economic savings. Realizing these savings requires that the DG either reliably coincides with local demand, or that it is controllable based on system needs. One way to accomplish this is with volt-VAR droop control, described in Section 14.4.3.

Third, to the extent that DG coincides with load, it tends to reduce the demand on transmission and distribution capacity such as conductors and transformers. Because this equipment has to be

sized to accommodate the peak, not average load, offsetting a small amount of load with DG during key hours can provide a significant relief for the T&D infrastructure. In areas with substantial load growth, local generation can serve as an alternative to adding T&D capacity. For example, the DG may make it possible to avoid or at least defer capital intensive upgrades such as reconductoring lines, or replacing a substation transformer with a larger model. Such benefits are highly situation specific.²⁶ The topic of load growth is timely in the context of strategic electrification for heating end uses and vehicle charging. The overall infrastructure cost of accommodating these new electric demands in the interest of decarbonization will depend on how much local generation can reliably mitigate net load, which in turn hinges on the degree to which both generation and load are controllable.

Fourth, generation in the distribution system impacts protection needs and coordination. This can be problematic, especially from the perspective of utility engineers faced with the challenge of adapting or redesigning circuit protection, to guarantee that any fault will be safely interrupted and isolated with minimal interruption to customers. In the standard radial distribution topology, power traditionally flows in only one direction, from substation to load, and the existing protection is likely designed and coordinated accordingly (see Section 7.5.2). DG introduces the possibility of power flowing in the other direction, and any section of line or piece of equipment therefore being energized from the side, on which there is no fuse or circuit breaker to protect it.

A related problem is known as *relay desensitization*. The crucial part of choosing appropriate settings for protection systems is to estimate the fault current under various scenarios. The presence of DG on a circuit means that its *fault current contribution* during the fault event has to be taken into account, which may alter the fault current sensed by a protective relay. If this fault current is less than anticipated, it could prevent the relay from recognizing and properly responding to a fault. For this reason, DG interconnections often require *protection studies*.

Beyond feeding power into a fault, the possibility of DG energizing an otherwise isolated circuit implies an electrocution risk for utility line workers²⁷; see Section 7.1.6 for more discussion of islanding. This protection issue has been somewhat controversial in the industry. While manufacturers and advocates of DG tend to feel confident that built-in features of the generation equipment called *anti-islanding protection* can guarantee that it will separate from the grid in the event of a disturbance, utilities often require additional disconnect switches accessible only to their crews. As discussed in Section 14.4.3 on inverter control, it is not trivial to distinguish between a minor disturbance and a grid outage based on locally observable quantities (voltage and frequency), and the distinction becomes more difficult as the penetration level of DG on a local circuit increases.

Another technical and controversial aspect of DG is the problem of availability and control. On the one hand, DG is generally *nondispatchable*, meaning that power system operators cannot call on it to provide power on demand or at specified times. The resource may be intermittently available (in the case of solar or wind) or controlled by the owner of the generation equipment; in either case, the utility or system operator has no control over operating schedules.

The lack of control is not necessarily problematic if DG behaves predictably and, ideally, coincides with local load so as to level out rather than amplify load peaks. This generally holds true

26 To be fair, if a large amount of DG is installed compared to the minimum daytime load, it is conceivable that not only will the distribution circuit become a net power exporter at the substation, but that this could require an increase of capacity.

27 Note that the electrocution hazard is associated with voltage level, not necessarily the amount of power provided by DG. For example, as long as a residential PV system remains connected to the grid by a 12-kV-to-120-V transformer, it energizes the primary side of that transformer with 12 kV regardless of the fact that it could never by itself sustain the loads in the neighborhood. Thus, even a very small DG installation could be deadly to line crews if not electrically isolated.

for solar power in areas of summer-peaking demand, although air conditioning load usually lags solar generation by a few hours. Customers may also have some local battery storage and use it to adapt their net generation and consumption to varying time-of-use rates, which would be designed to shave the load peaks.

Customer-owned generation of small or moderate size (e.g., up to tens of kW, depending on local regulations) is often connected behind the meter (BTM), without direct telemetry. In this case, the utility lacks not only control but any visibility of distributed resources: while they have records of how much capacity was permitted to interconnect, they cannot ascertain how much of this capacity is operational, or how much power it is generating at any given time. All that is visible to the utility or system operator is the net demand at the meter. Thus, any DG without telemetry has to be considered as a “negative load,” merging its uncertainty with that of the demand. The actual contribution of BTM solar generation has to be estimated with statistical tools based on load and weather forecasting, with increasingly refined algorithms.

Clearly, this approach is sensible when the relative amount of DG in the system is small, and becomes more problematic at increasing levels of system penetration. One problem with DG masking load from behind the meter is that system operators do not know exactly how much generation loss the overall system is exposed to—or conversely, how much load could be suddenly “unmasked”—in the event of a simultaneous DG trip due to a systemwide disturbance (for example, a frequency excursion that triggers the anti-islanding function on many thousands of inverters). This issue has motivated growing interest on the part of transmission-level operations to gain more insight into resources behind the substation.

The 2006 edition of this book stated that active centralized control over a large number of small DG units may be unrealistic due to the sheer information and communications volume. Information technology advances in the meantime have brought many control options into the realm of the possible. The extent to which it is desirable to coordinate and control DG centrally for optimal collective performance, rather than in a distributed or decentralized fashion, remains a strategic question. More local control generally sacrifices optimality for the practical benefit of being less dependent on communications (and thus less expensive and less vulnerable to disruptions), as well as requiring less computation and decision-making effort on the part of grid operators. It stands to reason that with growing penetration levels of DG, the motivation and incentives for active control of distributed resources in the service of the grid will increase—with many details to be negotiated about who exercises this control and how, or under what financial arrangements.

Last but not least among the systemic impacts of DG is that it would seem to reduce the grid’s vulnerability to disruptions, whether due to natural events or acts of sabotage, and increase the system’s *resilience*, or ability to recover from disturbances. In part, this is because the failure of any one small generation facility has less impact on the power system as a whole (besides making a less attractive target for attack). In part, too, DG implies less reliance on long-distance transmission links, and therefore would tend to reduce the scope of disruption caused by failures of individual transmission lines or substations. In principle, DG introduces the possibility of local self-sufficiency, which could dramatically reduce the societal impact of grid failures by making electricity available locally for vital applications during a crisis. However, the extent to which local areas or regions remain operable when the grid is compromised depends not just on local generation, but also on the connectivity details and on the information and control capabilities to orchestrate the available resources. For more on *microgrids*, see Section 15.4.

DG also has institutional implications for power systems at large. As generation is distributed geographically, it enters the jurisdiction of power distribution, as opposed to transmission, where power plants traditionally interface with the grid. For distribution engineers and operators, dealing

with generation is a fundamentally new responsibility—one, it might be noted, they cannot be expected to embrace with unqualified enthusiasm, as it entails new demands, complexities, and failure possibilities. Finally, there are important social and political dimensions of ownership of resources and generation assets.

In sum, an electric power system with significant amounts of renewable and DG represents a radical departure from the centralized and strictly hierarchical power system of the 20th century. This transition, well under way as of this writing, entails some important unanswered questions: while the grid's character is clearly changing, it is not yet obvious what, exactly, it is changing into, as many new technologies are compatible with different high-level grid design strategies. The societal implications extend far beyond electric power, as the use of distributed and renewable generation relates to diverse issues of resource scarcity, environment, and the politics of ownership.

While the social and political aspects are beyond the scope of the present discussion, we can identify some specific open questions about the grid's evolution in the face of DG. One has to do with the problem of *islanding*, or operating parts of the system while disconnected from others (see Section 7.1.5). On the one hand, the option of islanding allows extracting the full benefit from DG resources, that is, supplying power locally while the rest of the grid is unavailable. On the other hand, an interconnected system that could routinely operate with local or regional power islands—beyond individual customers with self-generation who are carefully isolated from the distribution system—represents yet another leap of technical and institutional transformation, with a host of controversial aspects including safety, liability, accounting, and control.

Another issue of practical consequence concerns the economic cost–benefit analysis of DG in relation to transmission and distribution infrastructure. While it was suggested earlier that DG offers systemic savings (e.g., on line losses or T&D investments), it is not obvious how this information can be used for strategic investment decisions when different corporate entities pay out the capital and realize the savings, respectively. Even if markets or regulation offer incentive mechanisms for DG, there remains the analytic problem of comparing and trading off what are qualitatively different investments. Limited experience exists in this area because, in vertically integrated and restructured market environments alike, generation and transmission planning have almost always been carried out by separate organizational entities, with separate accounting and separate responsibilities for justifying investments. This problem also relates to the economic evaluation of electric storage capacity within the context of the grid. The kind of systematic and integrated three-way comparison of generation, storage, and transmission capacity that one might imagine guiding the strategic development of power systems is not yet part of the standard analytic toolkit within the electricity sector.

Yet such an analytic framework becomes important as power systems include increasing contributions from intermittent, renewable resources. For example, we might want to compare the cost of electricity produced at two locations with different weather and different transmission distances to major loads, or that of a PV plant adjacent to a load with a fossil-fuel resource some distance away. Alternatively, consider a region where wind is the cheapest energy source available and supplies a substantial portion of local demand. Here, we might have to assess the maximum percentage of load that can be met by wind power before it becomes too unpredictable, and then determine to what extent the unpredictability ought to be compensated for by either introducing storage capacity, adding more expensive generation capacity from other sources, strengthening transmission ties to other regions, or simply oversizing the wind power plant.

In sum, the analytic challenge for sensible decision-making (whether by market mechanisms or some integrated planning entity) increases with the introduction of intermittent and distributed resources because of a sensitivity to time and geographic variables that 20th-century technology simply did not have.

15.3 Storage

Energy storage supports the balancing of power generation and demand. Storage can help guarantee adequate power supply during times of peak demand, when other resources are unavailable, or when demand ramps up too quickly for other resources to respond. This type of energy storage, relevant on time scales of minutes to hours, is usually what is meant by the plain term “storage” in the electric grid context. Other, less-common applications include very fast acting storage (on the scale of seconds or cycles) to manage power quality, and, at the other end of the spectrum, longer term storage that will become more relevant in the context of seasonally variable solar and wind energy.

Any type of energy storage will suffer some amount of *conversion losses* in the process of going to and from electricity, or *stand-by losses* that are sometimes called *self-discharge*. While specific technologies may perform much better or worse, a reasonable guess for the round-trip efficiency of most forms of electric energy storage, absent any specific information, is around 75%. The key idea is that the increased value of electricity at a later time will more than make up for energy losses on the order of tens of percent.

It is worth keeping in mind that in traditional electric grids, the role of energy storage has been played by fossil fuel inventories (such as piles of coal, tanks of oil, or natural gas compressed in pipelines), enabling dispatchable fossil-fuel power to operate as flexibly as expected. Physically speaking, stored energy in the legacy grid also includes the rotational kinetic energy of generators and motors, which is referred to as *inertia* rather than “storage” in practice, since it is a passive rather than an actively controlled feature.

A variety of active electric energy storage technologies exist today, ranging from innovative and precommercial to very well established. The answer to the question of where these various storage approaches have a solid business case is currently evolving, both on account of changes in the technologies themselves and in the economic valuation of storage capability in electric grids. Possible reasons for an increased value of storage include generation and transmission constraints leading to a high cost of accommodating demand peaks, as well as a growing contribution of nondispatchable generation such as solar and wind power.

The main take-away from this section is that there is a menu of technically feasible options that could be available at scale, if considered worth the price. Another important general observation is that different storage technologies are suitable for energy storage on different time scales. In the context of transitioning to a resource mix with high contributions from seasonally variable solar and wind power, long-term storage—that is, on the scale of months rather than hours—becomes increasingly important for ensuring resource adequacy, while fast-responding storage addresses a very different set of problems having to do with stability and control.

15.3.1 Hydroelectric Storage

In large electric grids, the energy storage medium of choice has long been pumped hydroelectric storage. While this technology has been around for over a century, it is constrained by available water and topography. A quick back-of-the-envelope calculation²⁸ shows that storing energy on the scale of megawatt-hours in the form of elevated water requires either a significant height difference or a large reservoir, and preferably both.

A *pumped hydro* storage plant requires a reversible turbine-generator and reservoirs uphill and downhill for water to be stored. The idea is to draw electric power from the grid when it is readily

²⁸ Using the formula $PE = mgh$ for mechanical potential energy, with mass m the product of volume and density of water, and using the gravitational acceleration $g \approx 9.8 \text{ m/s}^2$ and the comparative height h .

available and cheap, operating the turbine-generator as a motor to pump water uphill. This water is then available to flow downhill and generate power, by pushing the turbine as in a standard hydroelectric power plant, at a later time when electricity is scarce and expensive.

The size of the reservoir determines the total amount of energy that can be stored, and thus the time period for which stored power can be delivered. Although water reservoirs also provide seasonal storage (e.g., saving runoff from spring snowmelt for peak loads later in the summer), the typical application of pumped storage is for a diurnal cycle, pumping at night and generating during the day. The difference in the value of electricity between night and daytime peak hours easily compensates for the conversion losses: with the efficiencies of pumping and generating each in the neighborhood of 80%–90%, a round-trip efficiency of 75% is readily attained. Separate from the energy capacity, another relevant measure of a storage unit is its power capacity, or the rate at which it can absorb and redeliver the energy; this is a function of flow rates and machine ratings.

While the equipment is straightforward and the application convenient, the construction of pumped hydro storage units is constrained by topography, because large volumes of water and significant elevation gains are required to reach the scale of megawatt-hours. One prominent example of using pumped hydropower to great advantage is Switzerland, which serves as a storage bank of electric energy for the western European grid. Alpine elevations with plenty of water, centered geographically among dense loads, make this an ideal situation, not to mention a great business opportunity. In the United States, there are only few pumped hydro units, though their contributions are operationally significant.

15.3.2 Batteries

On the individual customer scale, aside from gas or diesel fuel for backup generators, electricity storage has become almost synonymous with batteries. Batteries can provide backup power during interruptions, but are also used just simply to store energy during times of abundant generation for later use. Battery technology has evolved rapidly and substantially in recent years, and larger batteries for microgrid as well as macro-grid applications are becoming increasingly relevant.²⁹

Batteries intrinsically work with direct current, so their use for a.c. systems always requires an inverter (see Section 14.4.1). The basic principle of energy storage in a battery is that an exothermic (energy-releasing) chemical reaction proceeds subject to the flow of ions through the battery and electrons through the external circuit. (Even though the electrons are tiny, they are unable to travel independently through the battery fluid.) When an electric load provides a path between the positive and negative terminals, the current can flow and the reaction proceed until the stock of chemical reactants is depleted. The chemical energy released in the reaction is equivalent to the work done by electrons in moving through the load, plus any heat losses. The battery is recharged by forcing the chemical reaction to reverse itself, which means forcing ions to flow backward through the battery by pushing a reverse current into the terminals with an electric power source. This recharging process consumes an amount of electric energy commensurate with the chemical energy to restore the original reactants, plus losses.

The performance of batteries is determined by chemical reactions on their internal electrode surfaces, including nonideal behaviors such as the irreversible buildup of chemical products that no longer participate in the reaction but just “gum up” the electrodes in some way. These processes

²⁹ The first, 2006 edition of this book stated that “on the scale of utility power systems, the amount of stored energy required to have any operational impact is so huge by comparison [to a single home] that batteries are simply not a realistic option.” With commercial grid-connected battery installations on the order of gigawatts as of 2022, that statement has become decisively outdated.

are affected by temperature, charge and discharge rates, the maximum and minimum voltage level reached, and the duration for which a battery is subjected to particular conditions. Different battery types vary greatly in terms of their sensitivity to these factors, and the extent to which they irreversibly degrade over the course of many charge/discharge cycles and operating stress.

Round-trip efficiencies for electrical energy into and out of a battery range from about 75% to 95%, depending on the technology and the circumstances of operation. Because the terminal voltage varies with the chemical conditions inside the battery over the course of the charging and discharging process, measuring the energy input and output for a charge/discharge cycle is not trivial, but requires integrating instantaneous power (current times voltage) over time.

The standard measure of battery charge and discharge is *ampère-hours* or *amp-hours* (Ah).³⁰ Amp-hours are straightforward to measure—we don't need to know the voltage at which each unit of charge was transferred—and have the advantage that electrical charge is, to a good approximation, conserved in the process of battery operation.³¹ The ratio of charge extracted during a full cycle to the charge added is called the *coulombic efficiency*, which can be as high as 99%. It is important to understand that when we speak of charge added or extracted, we don't mean the net electrical charge of the battery. Rather, we are referring to the charge transported by electrons through the battery terminals—both in and out—either in the direction where work is done on the electrons by an external source to force an endothermic reaction (charging), or in the direction where the chemically motivated electrons do work on the exterior (discharging).

Example

A lead-acid battery has a storage capacity of 80 Ah at 12 V. Neglecting losses, how many of these batteries would be needed to supply a residential load of 5 kW for 24 hours?

The energy demanded is $5 \text{ kW} \cdot 24 \text{ h} = 120 \text{ kWh}$. The energy stored in one battery is readily computed by combining volts with amp-hours, recalling that $1 \text{ V} \cdot 1 \text{ A} = 1 \text{ W}$. Thus, $80 \text{ Ah} \cdot 12 \text{ V} = 960 \text{ Wh}$. The number of batteries required is $120,000 \text{ Wh}/960 \text{ Wh} = 125$.

The *state of charge* (SOC) of a battery is related to the d.c. terminal voltage, but usually not in a straightforward, linear way over the range of battery operation. Because the terminal voltage is easily observed, it is often taken as a convenient, but imprecise, proxy for the SOC.³² By contrast, an accurate SOC estimate—not to mention an accurate forecast of how long the battery will last under the present conditions—requires more sophisticated analysis. Other advanced functionalities especially for large battery systems include intelligent charge control, discharge optimization, and balancing cells (which may be connected in series and parallel in large numbers) to equalize their SOC and voltage. Software for these processes may draw on physical measurement inputs (e.g., voltage, current, and temperature), operating history, time-varying electricity prices, and predictive algorithms. Optimization objectives may include supporting battery performance, prolonging battery life, and economically optimizing based on external factors. These sophisticated algorithms are typically contained within proprietary control software.

³⁰ Since ampères measure charge per unit time (see Section 1.1.5), an ampère-hour simply measures charge: $1 \text{ Ah} = 3600 \text{ coulombs}$, to be exact.

³¹ Total charge is exactly conserved in nature, but as some unwanted chemical reactions occur inside the battery, not all charge remains *available* for operation.

³² For example, a battery indicator that simply shows the terminal voltage will display a different value depending on the instantaneous current, and the battery level will appear to recover as soon as the load is reduced. This is a familiar phenomenon in many consumer products, but becomes more problematic with larger, more expensive battery installations or electric vehicles where accurate knowledge of the SOC can matter greatly.

Until the early 2000s, lead-acid batteries were by far most common, and mostly used in stand-alone (that is, not grid-connected) applications. Such lead-acid batteries for stationary use are similar to car batteries, but designed to be more tolerant of repeated deep discharge. As of the 2020s, after years of rapid advances in performance and cost, lithium-ion (Li-ion) battery technology has become the default choice for mobile applications, from small electronics to e-bikes to electric cars and trucks, as well as for moderately sized stationary applications such as powering a home. Though still expensive compared to the cost of electrical energy in itself, Li-ion batteries are available off-the-shelf for standard residential and commercial installations, especially paired with solar PV to allow for both grid and backup power. Numerous battery technologies and variants of different chemistries (including many options other than lithium) are quickly evolving, as is their respective significance for different applications.³³

Description of particular battery technologies and their comparative performance will not age well while this book is in print. One general observation is that, from the standpoint of electricity infrastructure, battery technologies are quite interchangeable. For perspective, the transition from gas stations (based on a vast petroleum refining and delivery infrastructure) to electric vehicle charging stations (supported by the electric grid) is an enormous societal undertaking. By contrast, shifting from one type of EV battery chemistry to another is a design change accomplished in a new production run for the next model year. In that sense, the shortcomings of specific battery technologies are likely not indicative of strategic limitations for battery storage in general.

It is helpful to consider the range of criteria by which competing storage technologies can be meaningfully compared and evaluated. Important local considerations related to battery chemistry include toxicity, corrosiveness, explosion hazard, and fire safety. Global considerations, especially in view of potentially scaling up production volumes by orders of magnitude, include scarcity of specific minerals as well as environmental impacts, health, and humanitarian concerns about mineral extraction practices (e.g., cobalt).

In terms of their functionality for various applications, crucial performance criteria for batteries—and energy storage technologies in general—include energy and density by weight, also known as *specific energy* (in units of Wh/kg), and *energy density* by volume (in units of watt-hours per liter, Wh/L), as well as the corresponding quantities of *specific power* (W/kg) and *power density* (W/L). Naturally, both weight and volume are important design constraints for mobile applications.³⁴ By contrast, bulk and weight are less of a concern for stationary applications, which includes almost all energy storage designed for connection to the electric grid. This lower performance bar may open opportunities for a greater variety of battery chemistries.³⁵

One important class of technologies for stationary applications is *flow batteries*. Rather than housing their entire chemical inventory in the same enclosure, flow batteries separate the functional reaction surfaces that make up the core of the battery from the supply of chemical reagents, which are kept in external containers as a fluid. The setup is similar to a fuel cell, except that the chemical process is reversible within the same device. While the surface area of the electrodes limits the

³³ For example, as of 2022, lithium iron phosphate (LFP) batteries are gaining market share relative to lithium manganese cobalt (LMC) and nickel cobalt aluminum (NCA) batteries, especially for stationary applications.

³⁴ The most stringent case of weight limitations is powering aircraft. Battery-powered experimental aircraft have been demonstrated, but this is a long way from commercial air travel. A quick back-of-the-envelope calculation shows that the energy density of aviation fuel (about 43 mJ or 12 kWh/kg) exceeds the energy density of today's commercial Li-ion batteries (about 270 Wh/kg) by a factor of about 40. Taking into account the conversion losses for fuel combustion engines, the advantage of chemical fuel shrinks to a factor of about 15—which remains a formidable challenge.

³⁵ It is also plausible for retired electric vehicle batteries, when they retain only a fraction of their original charge capacity, to be reused in “second-life” stationary applications.

power in watts, the volume of stored reagents determines the available energy in watt-hours, or the duration of discharge. Thus, the energy capacity can be increased simply by providing a bigger tank of reagents. Flow batteries can use many different chemistries, and exist at various levels of development and commercialization.

The proliferation of electric vehicles presents an opportunity to leverage battery storage resource using not only strategically timed charging, but active vehicle-to-grid (V2G) discharging in times of need. This requires a bidirectional inverter, as well as suitable economic incentives to compensate for any degradation of vehicle battery life.

15.3.3 Other Storage Technologies

15.3.3.1 Thermal Storage

A most elementary form of energy storage is *thermal storage*: keeping hot things hot, or cold things cold, for later use. The concept can be applied to the electric grid in innovative ways on either the supply or the demand side.

Any steam generation plant that deals with a heat exchange fluid can, in principle, set some of the hot fluid aside in a well-insulated tank for a few hours or more. With fossil fuels there was no point in doing this, since the chemical fuel itself sits on hand as a stored energy resource. Solar thermal power generation is an ideal application, where the stored heat (e.g., in the form of molten salt) can help compensate for variable sunshine, or delay power generation to meet electric demand during and after sunset.

On the load side, thermal energy storage is usually associated with pre-cooling at the commercial scale, to avoid peak demand charges on summer afternoons. Beyond just taking advantage of a building's thermal mass, chilling water or making ice increases the storage capacity for coolth. Note that the energy storage potential is greatly increased by taking advantage of the latent heat in a phase change. With suitable communication and control systems, it also becomes possible to recruit large numbers of smaller, *thermostatically controlled loads* (TCLs) such as refrigerators or water heaters as storage resources for the grid. Strategies may include fast response (on the scale of seconds) to signals from the grid operator or third party coordinator. The idea is to take advantage of thermal energy (warmth or coolth) that is already stored at the end user location, and varying the power demand in a way that has little or no impact on energy services as experienced by the customer (say, the ability to enjoy a hot shower or cold ice cream).

15.3.3.2 Compressed Air

At the utility scale, another possible energy storage medium is compressed air. In *compressed-air energy storage* (CAES), electric energy is used to operate pump motors that fill a confined space such as an underground cavern with air at high pressure. To retrieve the energy, the pumps are operated in reverse as generators. One interesting technical aspect is how to best manage the inevitable heating and cooling associated with gas compression and adiabatic expansion³⁶, respectively, in the context of plant design (e.g., by storing and recycling the heat). As with pumped water, a suitable storage location is the key constraint on siting large CAES facilities.

15.3.3.3 Flywheels

Flywheel energy storage is a modern implementation of an old concept, which is maturing into commercial use as of this writing. A flywheel stores rotational kinetic energy by way of a

³⁶ This is the familiar phenomenon of a bicycle tire getting hot when pumped up, and cold when the air is suddenly released.

fast-spinning disk or wheel, where the energy can readily be supplied and extracted using electric motor-generators. Unlike batteries, flywheels don't suffer aging as a function of charge-discharge cycles. The fundamental design challenge is to produce a wheel with high rotational inertia—which intrinsically requires it to be large and massive—capable of withstanding high rotational speeds (say, on the order of 10^4 rpm) while minimizing the risk and hazards due to fracture, and to scale up such systems at reasonable cost. Other interesting design aspects include minimizing friction losses (e.g., by magnetically suspending the wheel in an evacuated cavity), and advanced electric power conversion that allows for fast response (seconds or less).

15.3.3.4 SMES

Another fast-acting storage medium, but at lower technological readiness, is *superconducting magnetic energy storage* (SMES). It hinges on the ability to sustain extremely high currents within superconducting material, which are associated with a strong magnetic field. Since the “high-temperature” superconducting materials proven today require cooling by liquid nitrogen, this presents a challenge for SMES as a candidate for cost-effective grid energy storage compared to other alternatives in the near future.

15.3.3.5 Supercapacitors

Electricity can also be stored directly on capacitors or *supercapacitors*. As introduced in Section 3.3.3, a capacitor is a simple device designed to store electric charge. Its opposite plates effectively coax like charges (positive or negative) onto the same piece of conductor, where the work done to overcome the mutual repulsion of these charges represents stored potential energy. The familiar use of capacitors in power systems is for reactive power compensation (see Section 3.4.5), which physically represents an extremely short-term energy storage, where power is alternately absorbed and released within the duration of each a.c. cycle. In the storage context, large capacitors are used in uninterruptible power supply (UPS) systems to bridge the brief, sub-second gap during switching from one power source to another. However, expanding the energy storage capacity from cycles to seconds, minutes or even hours involves a vast increase in scale and cost. *Supercapacitors* are distinguished by having more than one pair of plates, so as to increase the energy storage capacity. Still, they are most suitable for short-term storage applications.

15.3.3.6 Hydrogen

At the other end of the spectrum, long-term or seasonal electrical energy storage can be accomplished by electrolyzing water into hydrogen and oxygen. Electricity is thus converted into a clean and convenient chemical fuel that can be stored in quantity, transported, and converted back to electricity by means of fuel cells.³⁷ Efficiencies of electrolysis range in the sixties and seventies of percent, making for a lower round-trip efficiency than other storage technologies.

Hydrogen from electrolysis using carbon-neutral electricity generation is known as “green hydrogen.” Note that the vast majority of commercially produced hydrogen as of this writing comes from reforming of natural gas (methane), sometimes called “gray hydrogen.”³⁸

Hydrogen conversion technologies are proven and commercially available, but there are significant economies of scale. As of this writing, it is neither easy nor cost-competitive for an individual enterprise to choose green hydrogen as a storage medium, absent a broad infrastructure committed

³⁷ See Section 15.2.1 on fuel cells. Hydrogen can also be used in a combustion engine, but fuel cells are generally much more efficient.

³⁸ Another term, “blue hydrogen,” signifies production from natural gas feedstock but with carbon capture. The actual climate impact of this approach is not obvious, considering upstream methane leaks.

to it. The key advantages of hydrogen are its suitability for mobile as well as stationary applications, the ease of extending the storage time without local constraints, and the ability to scale hydrogen conversion processes industrially to large volumes. This could ultimately provide the same level of convenience as fossil fuels, with eminently high energy density³⁹ and flexibility of use.

15.4 Microgrids

The term *microgrid* is applied to a range of configurations of electrical generation, storage and loads that can jointly operate as a *power island* in a controlled manner. A microgrid may be a stand-alone system, such as for a remote village, or it may have an intermittent connection to the main grid, such as a university campus or a hospital. The definition imposes no particular size range. In most contexts, one would assume a “microgrid” to include several buildings.

Microgrids may operate on alternating or direct current, with the former still more common due to compatibility with many devices. Instantaneous generation and load must be balanced regardless of scale. For an a.c. system, this means controlling frequency and voltage. In a d.c. system, only a single variable, voltage, indicates power balance and has to be managed.

Various voltage standards exist for both a.c. and d.c. The choice of voltage level for the microgrid distribution infrastructure would be commensurate with distance. For example, at several hundred volts it may be realistic to run wires for a hundred or so meters before losses and voltage drop become problematic (see Sections 1.4.2 and 1.3.3). One advantage of d.c. distribution infrastructure for solar and battery systems is that it avoids the need to invert and rectify power several times on the way into and out of the battery, saving a few percent of losses each time. This might be weighed against the consideration of particular needs for providing a.c. to specific loads, and perhaps using existing a.c. infrastructure.

The definition of a microgrid assumes no particular generation resource, nor that it is clean or renewable. Common resources include diesel or gas generators, and solar PV with battery storage.

One resource must be responsible for setting and maintaining the a.c. frequency. This can be a generator in isochronous mode (see Section 11.1.2), or a grid-forming inverter. Additional resources will operate on droop control or in grid-following mode. When connected to the macro-grid, all microgrid resources must switch into grid-following mode, since they cannot expect to control the frequency of the entire larger system.

The same applies to controlling voltage magnitude. A stand-alone generator would typically operate with an automatic voltage regulator (AVR), adjusting its reactive power output so as to maintain a constant bus voltage on the microgrid (see Section 10.4.2). While grid-connected, however, a small resource cannot expect to sustain a local voltage different from what the larger grid presents, and might burn itself out trying. The technology for these transitions is standard, but must be implemented correctly.

With a relatively small aggregation of loads, it is likely not economical to provide enough generation capacity to meet the maximum possible (noncoincident, see Section 6.4.2) demand, or enough storage to meet all energy needs for an islanding event of indefinite duration. Consequently, there must be a procedure for shedding load when necessary. Microgrid planning therefore involves assigning different priority levels, such as *critical* and *noncritical* loads.

³⁹ The energy density of H₂ by weight is 120 MJ/kg, nearly three times that of gasoline or diesel. The energy density by volume depends on the compression, with typical high-pressure tanks between 5000 and 10,000 psi. Other options for high-density storage include cryogenics or liquefaction, and material-based approaches where hydrogen atoms are absorbed within or adsorbed onto solids, such as metal hydrides.

A microgrid controller should have the executive ability to drive generation resources, curtail loads, and direct battery charge controllers when to charge and discharge. This may involve sophisticated optimization algorithms based on load, resource availability, prices, weather forecast, and reliability forecast (for example, based on advance knowledge or a probability estimate of a grid outage, and the estimated duration of an islanding event). With this information, the microgrid controller can choose the amount of each resource to recruit, and what SOC to maintain for batteries. In essence, it is a miniature version of the economic dispatch problem, except that transmission constraints are not an issue over the short distances of a microgrid.

Figure 15.2 illustrates a general concept of intermittently connected microgrids, where balanced clusters of resources (generation, storage, and load) may operate either independently as a power island or in grid-connected mode. In the context of today's standard practice and extant hardware, all distribution infrastructure within each shaded resource cluster in Figure 15.2 would have to be owned and managed by a single customer who interfaces (technically, legally and economically) with the electric utility at the point of common coupling (PCC).

As discussed in Section 7.1.6, power islands are not part of the *modus operandi* for traditional power systems. For a microgrid to disconnect from and reconnect to the main grid, there will be a specific *sequence of operations* agreed to by the microgrid operator and the utility. The most straightforward kind of transition is a *break-before-make* connection. Here, the load will experience a brief interruption (say, a few seconds) as the microgrid is isolated from the main grid and then restarts independently. When reconnecting, the loads will again be briefly interrupted before getting picked up by the main grid, and then the generation resources are added in grid-following mode. In this framework, there can never be any ambiguity as to which source is in control (local or main grid). The break-before-make principle also applies to homes with PV and battery storage, where the inverter toggles between grid-forming and grid-following mode while briefly disconnected.

When sensitive loads such as hospitals have an uninterruptible power supply (UPS), this comes with a *transfer switch* that can seamlessly transition the power supply from the main grid to the

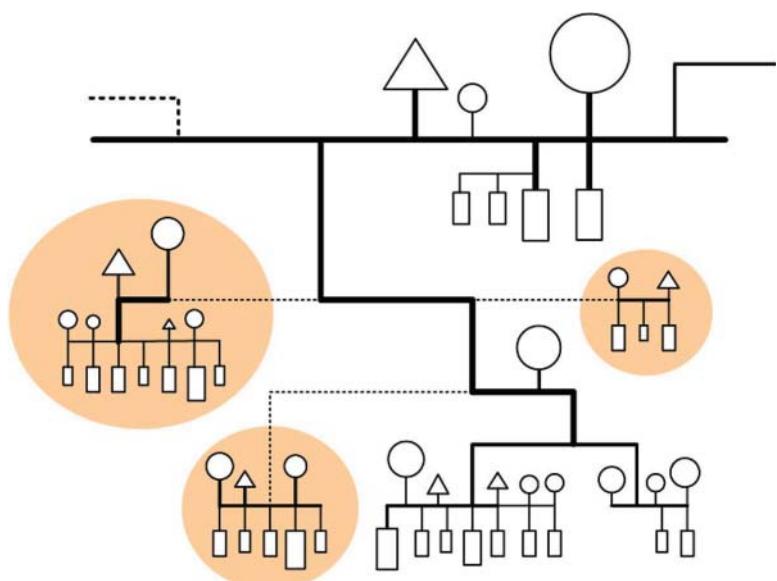


Figure 15.2 Microgrid concept, with generation (circles), storage (triangle), and loads (rectangles) at various scales. Dashed lines indicate optional connections.

backup generator and *vice versa*. Seamless in this context means that there is no electrical switching transient such as a voltage disturbance that would prompt any load to trip offline. This is accomplished with capacitors in the UPS that store enough energy for the load to ride through the transition. In UPS installations, the generators are typically intended strictly for backup needs, and never supply the main grid, which makes the interconnection less complicated.

The ideal combination of all worlds—generation that feeds the grid during normal operation, supports an island as needed, and transitions seamlessly between the two states—is possible to engineer, but not standard practice, since the value of avoiding a brief load interruption is not typically commensurate with the cost and effort this would require. The connection to the grid under load is called *paralleling* or *synchronizing*. It requires that frequency and voltage phase angle are matched closely enough across both sides of the switch at the moment of contact to avoid electrical transients or mechanical shocks. To perform this operation, it is necessary to have simultaneous control of the generation resource and the switch connecting the island to the main grid. In most microgrids, only the microgrid owner has visibility and control of the generation, while only the utility has control over the switch. Aside from the technical communication and control equipment, the issue of responsibility and liability during a seamless transition would have to be settled.

Another technical consideration for microgrids is that they may need a different protection scheme during islanded versus grid-connected operation. First, if there are several possible sources to energize a fault at any given location, the protection must be bi-directional. Local power electronic sources may not provide a high enough fault current to trip protective devices that rely on overcurrent, such as fuses and overcurrent relays (see Section 7.5.2). Some protective devices use phase imbalance to recognize a fault, but this may be more difficult to discern if loads on the three phases of the microgrid are already poorly balanced.⁴⁰

As of the early 2020s, there is growing interest in microgrids in response to declining trends of grid reliability associated with extreme weather. Some electric utilities are exploring innovative service models, such as supplying home-scale or community-scale batteries in lieu of reinforcing transmission and distribution infrastructure. The idea is that under some circumstances, building and operating a microgrid may be a less expensive way even for a traditional utility to provide reliable service to its customers.

While the state of modular generation, storage, and control technologies has advanced to make microgrids technically feasible at just about any location and scale, some important challenges to microgrid expansion in the United States are legal and regulatory in nature. For example, regulations generally preclude electric utility customers from sharing or transacting power across property lines (lest they form their own regulated utility). Microgrids that span multiple buildings or facilities are almost always owned and operated by a single entity, such as a corporation or university.

Conceivably, a focus on societal resilience and emergency preparedness could motivate new regulations to create opportunities for multi-customer community microgrids that can function with available local solar and storage resources, perhaps on an ad hoc basis during major grid outages. This will require some entity to be responsible for managing not just generation and load balance, but power quality and protection under all operating configurations. What would have seemed an insurmountable coordination challenge decades ago is readily envisioned today, at least on an engineering basis, with modern information technology, ubiquitous communications, and plug-and-play standardization.

⁴⁰ This is a good opportunity to reiterate that the aim of this book is to offer qualitative understanding of the issues, rather than the power engineering skill set to actually design and build solutions. Power system protection is a particularly esoteric area of expertise, yet critical for safety.

16

Making the System Work

The electric grid has been described as a system that works in practice, not in theory.¹ It is often referred to as a *complex* system. This makes perfect sense colloquially, as the notion of a vast array of metal hardware and moving machinery resonates with the standard definition of complex as “consisting of interconnected or interwoven parts” or “involved or intricate, as in structure; complicated.”² Indeed, a system’s complexity can be defined as the product of the number of its components, their diversity, and the *tightness of coupling* among them, meaning the strength and immediacy of their interactions. Throughout this text, we have seen many examples of such tightly coupled behavior.

The components themselves are curiously diverse, with a juxtaposition of old and new technology. State-of-the-art inverters and finely tuned gas turbines synchronize with hydroelectric units dating back to the early 20th century. The hardware of electric power systems is some of the oldest industrial machinery still in general use today. The essential tasks of transmission and distribution infrastructure remain unchanged: to connect pieces of conducting metal into electric circuits, to step voltage up and down, and to safely protect circuits in case of problems. While advances have been made in the design and materials used in transformers, capacitors, and circuit breakers, the basic functions of these power system components are remarkably simple. Likewise, synchronous generators have seen mostly subtle refinements, not fundamental changes of design. Many of today’s power system components could be replaced with well-preserved models from the early 1900s without compromising their role in the grid.

The simplicity of many of the individual pieces of hardware stands in remarkable contrast to the complexity of the interconnected system. As a whole, the grid presents itself as a difficult-to-predict entity with subtle behaviors and unexpected sensitivities. A pragmatic definition of a complex system is one where no individual can, at any one time, understand the entire thing.³

What might seem “complicated” or hard to understand to the layperson is in fact not fully understandable even to the expert. There is just so much of it that, by necessity, every power system professional has a limited scope of expertise and responsibility. Keeping track of the interactions of all these domains—the performance of the system as a whole—is by necessity a team process. This is true regardless of whether the terms “individual,” “expert,” and “professional” are taken to represent human beings, computer systems, or artificial intelligence. No single entity can

1 Todd R. LaPorte and Paula M. Consolini, “Working in Practice, But Not in Theory: Theoretical Challenges of ‘High-Reliability Organizations,’ *Journal of Public Administration and Research Theory*, 1(1), 19–48, 1991.

2 *American Heritage Dictionary* (Boston, MA: Houghton Mifflin Company, 10th edition, 1981).

3 This definition follows Barbara Czarniawska, who writes that “an organization becomes complex when no one can sensibly and comprehensibly account for all of it.” [B. Czarniawska-Joerges, *Exploring Complex Organizations* (Newbury Park, CA: Sage Publications, 1992).]

simultaneously monitor, control, and troubleshoot every generator, every load, every piece of conductor in between, and every possible external disturbance—any one of which has the inherent ability to affect every other system component almost instantaneously, and sometimes severely. Such complexity makes it possible for a system to surprise both its operators and its designers.

16.1 Time Scales for Operation and Control

The central challenge in the operation of electric power systems is often cited: electricity must be generated in the exact moment that it is consumed. To move electric power through a grid is to obey the law of energy conservation: what goes in must come out. Unlike a natural gas pipeline that can accommodate a variation in gas pressure and thus serve simultaneously as conduit and storage reservoir, a transmission line cannot store electricity. Power systems may include storage resources like pumped hydroelectric plants or batteries to smooth out diurnal or seasonal variations in demand and supply (see Section 15.3). But while a facility that alternately absorbs and releases power can help with saving up capacity for when it is needed, it does not circumvent the fundamental problem of coordinating generation and load in real time.

If we look carefully, we find that some physical energy storage is in fact provided by the standard components of a power system; there just isn't very much. In reality, if there were exactly zero energy storage, it would be literally impossible to operate a grid: every control action would need to be perfectly precise and instantaneous, or the alternating voltage and current would collapse. What we really mean when we say that generation and load have to be exactly balanced is that they have to be equalized *within the time scale* permitted by the system's capacity to buffer the discrepancies, that is, to store and release energy, and to adapt to new operating conditions.

The intrinsic short-term energy storage capacity in a conventional power system resides within large rotating machinery. As discussed in Sections 13.4 and 11.1.1, generators stabilize system frequency through their rotational inertia, absorbing and releasing kinetic energy in response to changes in the electric load. This process takes place in a fraction of a second, meaning that the amount of energy involved is comparatively small, but it represents the first line of defense against power imbalances in the grid. With increasing contributions from inverter-based generation that do not have intrinsic physical inertia, active provisions must be made for a stabilizing response to fluctuations. This is possible because power electronics are internally controlled on a much faster time scale than 60 Hz.

On the load side there is also some degree of flexibility because power consumption is not precisely fixed. Although a load's power demand is usually modeled as an externally given, independent variable that determines everything else, it does in fact afford a modicum of an inherent stabilizing response to power imbalances. Specifically, if real power into the grid is less than real power out, the a.c. frequency will drop, as stored kinetic energy from the generators is being used (Chapter 11). But at a lower frequency, some loads (motors, in particular) will also consume a bit less power. Although the system is not intended to be operated in such a state of underfrequency, the point is that on a very fine scale, load is not perfectly rigid. This applies not only to frequency, but also to voltage (see Section 6.3).

These small degrees of flexibility notwithstanding, the law of energy conservation is strictly enforced by nature. What little wiggle room there is can be quickly used up. It remains true, then, that the prime directive for power system designers and operators is to balance generation and load at every instant. This balancing act occurs on multiple levels, with control methods appropriate to each time scale.

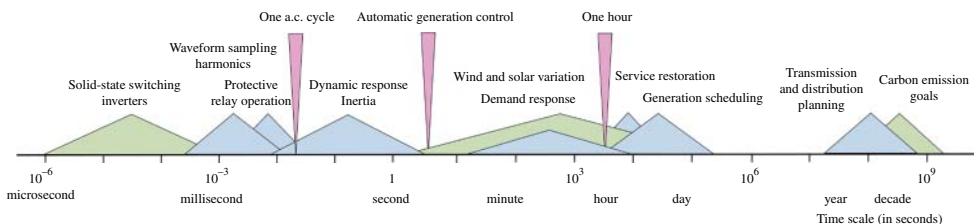


Figure 16.1 Time scales in electric grid operation.

Figure 16.1 places various different aspects of grid operations on a logarithmic time scale. The units are seconds, and the scale of relevant phenomena ranges from 10^{-6} (μs) to 10^9 (gigaseconds), better known as decades. Reference markers on the graph indicate one a.c. cycle, four seconds as a standard interval for automatic generation control (AGC), and one hour as a standard interval for market transactions.

16.1.1 Fast Response

The most sensitive aspects of maintaining equilibrium in a power system happen within fractions of a second. To the human observer, these events appear instantaneous, and their speed certainly demands automatic responses of the technical system components.

Below one cycle, voltage, current, and power are not defined in conventional terms of root-mean-square values, only as instantaneous quantities, and “frequency” is not defined. Phenomena where this perspective applies include faults, harmonics, and other transient disturbances—in other words, nonideal behaviors of the power system.

Historically, the only type of intervention performed on such a fast scale was protection (Section 7.5). Protection means that in the event of a fault, an accidental contact, or a short circuit on any system component, the current flow is interrupted automatically and as soon as possible or practical in order to prevent harm to people or equipment. While some devices such as fuses, are triggered by heat and may require several seconds of high current to melt, more sensitive mechanical relays can interrupt a current within one or several cycles. Since time is of the essence, circuit protection generally occurs automatically and has relied on local information, without supervision or intelligent intervention that would introduce latency.

Today, many types of power electronic devices interact with the grid, including loads, inverter-based generation, high voltage direct current (HVDC) converters, and flexible a.c. transmission systems (FACTS). The internal switching frequencies of power electronics are often in the tens of kilohertz. Control settings can be programmed to shape the behavior of these devices *within* each cycle, as with a fine sculpting tool.

The regulation of synchronous grid frequency begins on the time scale of a cycle. As discussed in Section 11.1.2, the basis of frequency response is the passive negative feedback effect built into the physics of synchronous machines: if the generator speeds up, the torque holding it back increases as the result of an increasing magnetic field; if the generator slows down, the restraining torque decreases. Here, the term “passive” means that this effect requires no intervention on the part of any human or machine to take place; it is intrinsic to the device and guaranteed by physics.

Figure 16.1 summarizes the above as “dynamic response, inertia.” This also includes unintended consequences such as wide-area oscillations and other stability issues. Both angle and voltage stability (Section 13.4) reflect physical processes—specifically, exchanges of energy in the form of electric and magnetic fields—which occur on the time scale of a cycle. The implication is that

once a power system reaches instability, it is probably too late to take any corrective action: the actual situation can evolve much more quickly than the latency in receiving and processing information, or doing something about it. For this reason, an important aspect of real-time operations is to keep a careful eye on the state of the system, try to anticipate if there is any chance that a nascent imbalance might deteriorate, and take preventive actions long before there is any obvious sign of trouble.

16.1.2 Real-Time Operation

Most of what would be described as “real-time operation” of the electric grid occurs on time scales of seconds to minutes. This includes both automated processes, mostly for routine tasks, and human actions, especially when nonobvious decisions must be made. Human intervention in real time may be called for at individual generation units, at the system operator level where systemwide generation and load are balanced, and in transmission and distribution switching.

While power plants are generally designed to provide constant power output without human intervention, operator actions are often required during start-up or shutdown, and sometimes to implement changes in output. Starting up a steam generation unit is a demanding procedure that involves coordinating numerous pumps, valves, flow rates, pressures, and temperatures throughout the plant. From a cold start, it takes hours to bring a large steam plant into its hot operating equilibrium, at which point it can be electrically connected or *paralleled* with the grid. Once steady output is reached, operators focus on monitoring automated processes until they need to initiate major changes in plant output, or coordinate clearances for equipment maintenance while the plant is online.

It is possible but not part of standard procedure for human operators to manually match a generation unit’s output with load when the load variations exceed the normal range of the governor system. An experienced operator can do this by watching frequency and voltage levels and adjusting valve settings accordingly. Such skill might be called upon in emergency situations where a plant is supplying a power island or part of a severely disrupted grid.⁴

At the level of system operator or dispatcher, the goal is to arrange for the correct amount of real and reactive power actually demanded by the system, as opposed to the amount previously estimated and contracted for. While power might be scheduled administratively on an hourly or 15-minute basis, these schedules cannot be physically accurate for several reasons: first, load depends on consumer behavior and cannot be known with certainty ahead of time; second, even if the forecast is generally correct, the load will still vary throughout the time interval; and third, generators may not actually produce what they promised. Another small contribution to the uncertainty comes from line losses, which depend on the geographical distribution of generation and load. As a result, neither supply nor demand, nor any discrepancies between them, can really be known until the moment they are measured empirically.

The process of making continual adjustments to match generation and load, as well as reconciling power imports and exports to neighboring areas, is detailed in Section 11.1 on *load frequency control*. At the level of a regional balancing authority, AGC instructions for individual generators are computed on a fully automated basis. The AGC signal marks the lower bound on the time scale for remotely communicated control commands.

⁴ For example, operators at Pittsburg Power Plant in California recalled keeping their units online “by the seat of their pants” after the 1989 Loma Prieta earthquake (personal communication). In a nearly complete communications blackout, their only guide as to how much load might still be connected out there were their local frequency and voltage measurements.

If the automated process fails to produce an acceptable outcome—that is, the departure of actual from anticipated system conditions exceeds the range of automatic controls—human operators may intervene at various levels. For example, the system operator or balancing authority may have to call upon power plant operators to take action. Such calls are often made by telephone, and in critical situations their success may depend on personal rapport between individuals.

A more drastic intervention on the part of a system operator or utility is to *shed load*. This may mean selectively disconnecting large customers with specific contracts and remunerating them for being *interruptible loads*, or local groups of customers in *rotating outage blocks* that are assigned to spread the burden of power shortages evenly, with outage blocks taking turns of an hour or so of blackout. The disconnection is physically carried out at the distribution level.⁵

Another example of real-time intervention is transmission and distribution switching to reconfigure the network for maintenance and restoration purposes or to preempt local problems such as overloading a particular circuit. These processes essentially involve opening and closing specific switches or circuit breakers according to a carefully mapped sequence, so as to isolate and connect parts of the grid without overloading any component and while maintaining appropriate circuit protection.

16.1.3 Scheduling

Whereas real-time operation is dominated by technical considerations, the scheduling context emphasizes optimization around economic criteria: specifically, which generation units to operate when and at what power level so as to minimize overall cost. This optimization will be *constrained* by technical parameters such as transmission line loading limits and security, but economics are front and center. While economic and technical objectives resided under the same administrative roof in the traditional, vertically integrated utility setting, competitive market environments aim to separate the function of dispatching resources from the ownership of these assets so as to prevent conflicts of interest. For example, a system operator should have no incentive to preferentially call upon certain generating units, perhaps under the guise of technical necessity—thus the term *independent system operator* (ISO). ISOs interface with the owners of generation assets or intermediaries, such as *scheduling coordinators* who submit bids into the market, and with transmission system owners (often utilities). In some countries, the functions of system operations, planning, and market facilitation are further disaggregated—for example, into a *transmission system operator* (TSO) responsible for the physical operation the grid and a *Power Exchange* to facilitate markets. This chapter aims to be agnostic to the business entities and regulatory environment, focusing on technical and pragmatic aspects.

The process of scheduling generation units days and hours in advance to be ready to produce a certain amount of power at a certain future time is called *unit commitment*. The idea is to match the forecast system load at minimum overall cost, given the particular production costs, locations, and operating constraints (such as rated power, ramp rate in megawatts per minute, or time required to start up) of each unit. Algorithms for this type of optimization with large numbers of variables are the subject of continuing research and development, as computing tools and market participation rules evolve.

⁵ The February, 2021 power outages in Texas during an extreme cold weather event are an instructive case study of a vast mismatch between load and available generation, which required drastic load shedding on an unprecedented scale. The problems with this event included not only failures of forecasting, weatherization and natural gas supply, but also a lacking granularity in selectively shedding customer loads. One analysis can be found at <https://cgmf.org/blog-entry/435/REPORT-%7C-Never-Again-How-to-prevent-another-major-Texas-electricity-failure.html> (accessed September 2021).

In the “Old World” of vertically integrated utilities that own and operate all the generation and transmission assets within their service territory, the decision of which generator contributes how much and when is made in a central scheduling process by means of an *economic dispatch* algorithm. An elementary, classic textbook approach to *economic dispatch* is presented in Section 11.2. The term *dispatch* refers to specific instructions for the amount of power requested from a generator at particular time. Although there is no formal definition that applies across electric power systems worldwide, dispatch is generally understood to occur on a shorter time scale than unit commitment, which identifies generators expected online and ready for dispatch instructions.

In competitive markets, unit commitment can be determined through some combination of *bilateral contracts* between individual parties, and a *power pool* that serves as a clearinghouse for power bought and sold. The detailed institutional arrangements vary among countries and jurisdictions and are far beyond the scope of this text.

From the technical perspective, what is important is that some entity keeps track of megawatts bought and sold, and a system operator (which may be the same or a different organizational entity) keeps track of the balance of generation and load. There may be multiple such entities within a geographic area. They might solicit generation through some form of auction, calling upon the lowest bidders to generate during each hour, where the auction could include *day-ahead*, *hour-ahead*, and shorter-term markets. In concept, such an auction should produce a picture similar to the one depicted in Figure 11.8, with each part of the annual *load duration curve* filled in by the least expensive generation available at that hour.

While day-ahead and longer-term planning or contracting are thus intended to create an economically optimal schedule, some modifications and adjustments are always necessary on short notice to accommodate factors like changing unit availability or maximum ramp rates. In an auction, these factors can in principle be accounted for by each generator bidding its available power specifically for each hour. In addition, though, the grid relies on generators to be responsive to real-time changes in demand, or any other factors (say, an outage somewhere) that affect the overall power balance. Such responsiveness may be specifically contracted for in ancillary services such as spinning reserve⁶, load following or frequency regulation, where generating units are remunerated for being “on call” to respond instantly to the grid’s needs.

Another component of ancillary services is the provision of reactive power or MVAR (see Section 3.4.2). It usually does not cost a power plant very much to provide reactive power (since, to a first-order approximation, it uses no fuel energy), and one might therefore expect MVAR allocation to be a somewhat casual, *ad hoc* process. In reality, it is necessary to schedule reactive contributions from specific generators ahead of time in order to achieve an optimization similar to the way real power is allocated—in this case, maintaining a certain voltage profile across the grid while minimizing overall line losses. In a competitive market, these services make for additional business opportunities.

Given a proposed generation schedule, the system operator must ascertain that it does not violate any technical operating constraints, such as the maximum load on a transmission line or an $N-1$ security criterion (Section 13.3). This is known as *security-constrained dispatch*. The system operator requires schedules to be revised if a violation can be identified in advance or makes adjustments in real time as needed. With the large number of generation and load variables, their intrinsic uncertainty, and the difficulty of controlling them directly, it is not uncommon for what was conceived as a scheduling task to become effectively a real-time operations task, presenting a challenge to system operators that is not to be underestimated.

6 The “spinning” property of the generation reserve means that the turbine and rotor are literally spinning at synchronous speed and able to pick up megawatt load without having to warm up or even accelerate.

Any crisis situation in a power system means that the time horizon for decisions shortens, and the focus in operation shifts accordingly from economic optimization to meeting the physical requirements of the grid at any cost. This shift is implicit in the shared assumption by all participants that the electric grid as a whole is to be kept functional at all times and under any circumstances if humanly possible. In economic terms, the demand for generation and ancillary grid services becomes *inelastic* at the edge of the grid's operating envelope. The reality is that the ideal of cost minimization in scheduling always remains subject to being outdone by events in real time—regardless of the institutional or regulatory framework.

16.1.4 Planning

Hourly and daily generation scheduling as well as the real-time operation of power systems take place within a set of boundary conditions that include extant generation, transmission, and distribution capacity and loads. These boundary conditions are addressed in the realm of planning on a time scale of years.

Historically, the planning process has been driven almost entirely by load forecasts.⁷ The premise of the traditional regulated monopoly framework is that all demand within a given territory is to be served by means of prudent investments sufficient to satisfy this demand. Planning then comes down to estimating load growth in megawatts, locally and systemwide, and accommodating this growth with appropriate upgrades in transmission and distribution hardware, new construction of generation units, or securing of electricity imports.

In view of consistently increasing electric demand and guaranteed future revenues under the historical U.S. regulatory framework, it made sense for utilities to adopt a long-term horizon in their planning. This meant getting ahead of load growth with oversized transmission and distribution (T&D) capacity that would come to be utilized over the years (the analogy of children's clothing comes to mind). Proactive investment in T&D infrastructure was understood to be justified both by the cost structure of upgrades and by the operational benefits of a robust grid. As of the early 2000s, however, much of the slack built into T&D capacity in the U.S. has been taken up. Aside from the economics, siting and permitting is a key constraint for building new transmission lines, with long project lead times and significant risk of cancellation.

Generation planning, though carried out by different entities and with many different constraints, shares some of the same challenges. For both technical and institutional reasons, the lead time for generation projects is considerable, sometimes with more than a decade passing between a unit's inception and its operation. In the regulated U.S. industry, it fell to utilities to maintain a generation reserve margin, traditionally 20% above peak demand. Therefore, unit construction plans were typically based on conservative demand projections for 10 or 20 years out. This approach made sense until the 1980s, when electric demand growth fell behind projections and many U.S. utilities suddenly found themselves with excess capacity.

In the competitive market environment, there is no simple recipe for ensuring strategic and timely capacity investment by independent firms. For example, the design of the deregulated California market in the 1990s assumed plenty of excess generation capacity to be in place and therefore failed to carefully consider the system's behavior under a hypothetical generation shortage. After

⁷ The term "load forecast" is generally viewed within the industry as a purely technical parameter determined by population growth and consumption levels as independent variables. Of course, it can also be understood as the expression of a more complicated social dynamic in which the industry itself plays a role—say, through advertising campaigns or incentive programs. The latter sort of "strategic" planning tends to be organizationally separate and conceptually remote from the "tactical" planning by power system engineers.

a dramatic shortage in late 2000 resulted in extreme wholesale price spikes (bankrupting the California Power Exchange merely three years into its existence) and subsequent market disruptions culminated in rotating outages for customers, the crisis was later determined to have been largely manufactured through gaming behavior by market participants.⁸

In theory, a competitive market ought to provide incentives not only for short-term production but long-term investment, including generation and power delivery. How such investment signals will occur in practice, and how closely the results will match society's expectations, is anything but clear. Even with parties ready to invest, the interconnection queue may be slowed by regulatory and environmental approvals.

In the context of decarbonizing the grid, it is important to recognize the interdependence between generation and transmission projects. Solar and wind farms make no sense to build in locations with great resource but inadequate transmission capacity to load centers, while transmission projects make no sense to build without a "there" there. Regardless of whether the entity investing in transmission is a utility or some independent third party, analyzing the business case for new transmission lines based on their quantifiable marginal value to the grid and any associated revenue may underestimate their systemic and strategic value. To meet aggressive policy goals and schedules for a sweeping transition to renewable resources, none of the planning and investment strategies from the past may prove adequate.

16.2 Measurement and Data

16.2.1 Historical Notes

By 21st-century standards, electric grids are operated with surprisingly few direct physical measurements, and much less systemwide observability than one might imagine. This makes sense in that power systems evolved in an era of great information scarcity compared to today. Considering that in the late 1800s, few locations were even connected by telephone, it was necessary to carry out all major operational functions—including controlling frequency, voltage, and switching operations—based on purely local information that could be readily observed.

Not all this information has to come from direct electrical measurements. For example, real power balance in the grid is accomplished by generator droop control (Section 11.1.2), where the key quantity, a.c. frequency, is observed by way of a generator's mechanical rotational speed. The open or closed status of a traditional switch or circuit breaker could be ascertained by visual inspection. The brightness of an incandescent lamp would give at least an approximate sense of voltage. But to ensure the physical operating parameters are kept within the design range and limits, and to track energy and power transactions, some electrical instrumentation is required.

The two essential quantities are voltage and current. Voltage is measured to ascertain whether the system is operating within its nominal range, usually $\pm 5\%$ or $\pm 10\%$. Voltage measurements are also relevant for dispatching reactive power. Traditionally, this is done only at relatively few locations in the grid, where direct control can be exercised: at generators, to inform the automatic voltage regulator (Section 10.4.2), and at distribution substations, to inform transformer tap changes (Section 7.4.1). Voltage is also needed for an accurate power measurement, but the key variable for measuring power and energy transfer is current. Traditionally, this would be anywhere that energy

⁸ The most famous offender was Enron Corporation. A synopsis of the 2000–2001 California crisis can be found at https://en.wikipedia.org/wiki/2000%E2%80%932001_California_electricity_crisis.

is transacted: at a generator bus, and at a customer meter. Current is also measured to monitor the load on major pieces of equipment, such as transmission lines and substation transformers. Historically, these would be analog measurements that a person has to go read on location.

Around the 1970s, supervisory control and data acquisition (SCADA) systems were introduced, initially at power plants and in high-voltage transmission, and later in distribution systems. Like in other types of industrial facilities, SCADA provides readings from instrumented equipment throughout the plant and allows operators to actuate components from a control room—instead of having to, say, traverse the turbine building, several flights of stairs and a maze of pipes to read a dial or turn a valve.

In transmission and distribution, SCADA includes the remote operation of equipment such as switches and circuit breakers to reconfigure the system topology. Sensing and control nodes in the field, known as remote terminal units (RTUs), are connected to a staffed control room by some channel deemed sufficiently reliable and secure. Physical communication layers may include dedicated telephone landlines, wireless signals such as microwave radio, or power line carrier signals.⁹ Previously, operators relied on field personnel to physically travel to each location in order to report equipment status or undertake switching actions; many rural distribution systems are still operated in this manner. SCADA sensors typically report a measurement value every few seconds. As of this writing, higher-resolution and time-synchronized types of sensing and measurement (see Section 16.2.4) are increasingly deployed.

Without some communications system overlaid on the power distribution infrastructure, operators have no way of determining whether customers are connected or whether a problem has occurred somewhere in the field. Unless the problem has proliferated to the level of a manned substation (say, by tripping a circuit breaker), nobody but the customer might notice a local disturbance, as the grid offers no intrinsic way of measuring from one location what is happening in another. To reduce both the hazards of undetected faults and the duration of service interruptions, some utilities installed outage report systems that identify trouble spots by correlating customers' telephone numbers with their locations on the grid (assuming that at least some fraction of customers will call to complain).

More recently, *advanced metering infrastructure* (AMI) provides automatic outage status reporting to the utility. Digital “smart meters” with wireless telecommunications began to replace traditional analog kilowatt-hour meters in the early 2000s. Traditional rotary meters showed only cumulative energy consumption and had to be read visually by a human meter reader. Having been designed more for billing purposes than to support grid operations, smart meters typically report only power outage status in real time, and power data in crude time intervals with some delay (e.g., 15-minute kilowatt demand data, sent to the utility every few hours).

Yet empirical measurements are fundamental to support “smart grids,” or power systems with the ability to observe and control processes at finer resolution in both space and time. Advancements of this type rely on detailed offline data to inform the development and tuning of algorithms, and appropriate real-time data to inform decisions by humans or machines. Increasingly detailed visibility of the electric grid along with informative analytics using large data sets is an area of active research and development.

One challenge for the industry—at least in the United States, to date—has been transitioning to a more sweeping and visionary approach to grid monitoring, from the traditional approach

⁹ In communication by power line carrier, the information signal is simply superimposed onto the 60-Hz power signal along the conductor. Having a smaller amplitude but much higher frequency, this signal is readily isolated by equipment on the receiving end; nevertheless, the equipment must be designed to safely interface with high voltage.

of collecting limited measurements and data for specific purposes. In a context where information is generally expensive, investments in sensing and data communication had to be justified with specific use cases for any given installation. Vendors often sell packages of sensor hardware, communications, and associated software to utilities for solving specific problems. But this type of proprietary technology does not necessarily interface with other data the utility might be collecting, leading to information “silos.”

For example, information collected by smart meters may be accessible to a utility’s billing department, but not operations. The meters would typically measure local voltage, but not communicate it. Perhaps the customer meter can be queried for voltage, but there will be no platform ready to integrate that information with SCADA data from the substation.

As instrumentation and data handling technologies become less expensive, and as our digital culture has higher expectations for interoperability of information technologies, it increasingly makes sense to record detailed electric grid measurements for general purposes. In a world where “terabyte” or even “petabyte” is no longer an intimidating unit, it becomes realistic to store, query, and cross-reference detailed measurements simply to gain overall awareness and to answer a plethora of possible questions about how the grid is working—including ones not envisioned at the time the data-collection infrastructure was conceived.

There is no shortage of novel analytic and operational questions in power systems, especially given the growing importance of inverter-based generation resources and other power electronics. The tools and approximation techniques introduced in this book—parts of the standard traditional power engineering curriculum—are largely based on assumptions about rotating machines, sinusoidal waveforms, and simple analog devices. The shift to digitally controlled power electronics on a large scale means that traditional assumptions about system behavior may become less apt, particularly in areas such as stability analysis that involve the interaction of many components. Heuristics or rules of thumb on which power system operators and engineers have long relied could suddenly fail. To maintain confidence in one’s understanding of the grid as a complex system, data-intensive tools based on direct measurements will likely become increasingly important.

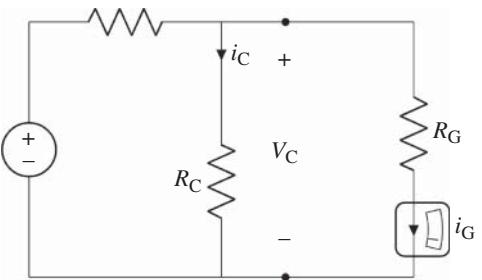
Enabling technologies for such a transition include not only low-cost sensor hardware, but flexible and secure databases that support a variety of user needs. With these technologies evolving, trends point toward increasing spatial and temporal resolution of grid measurements, and a wide open space of opportunities for new analytics based on *empirical* data to complement the classical power system analysis tools discussed in this book.

16.2.2 Physical Measurements

It is worth considering how voltage and current measurements are physically made. In order to observe an electrical circuit in the conventional sense, the instrument must become part of that circuit. Important exceptions to that rule include optical sensors and those that detect electric or magnetic fields surrounding the conductors. Such remote sensors have the distinct advantage that they do not need to be in direct physical contact with hazardous high-voltage equipment.

The classical analog device for measuring either current or voltage is a *galvanometer*, implemented in diverse configurations. In the galvanometer, current traveling through a coil creates a magnetic field and force, which pulls the pointer in opposition to a mechanical spring force and deflects it to the corresponding position on the dial. The galvanometer is introduced into a circuit in combination with suitably sized resistances to scale the amount of voltage and current to a level compatible with the instrument (say, single-digit volts or amps). By applying Ohm’s law to the series and parallel components, the measurement of current through the instrument itself can be

Figure 16.2 Making a voltage measurement on a parallel branch.



converted to read out the current through or voltage across the external circuit branch of interest. Minimizing the amount of power dissipated by the instrument (in the interest of practicality and safety) also minimizes its distortion effect on the rest of the circuit (in the interest of measurement accuracy).

Example

To measure the voltage v_C across a circuit element or branch, a galvanometer along with a series resistance R_G is connected to the points of interest so as to create an additional parallel branch to the original circuit, as shown in Figure 16.2. Choosing a large value of $R_G \gg R_C$ results in a small current $i_G \ll i_C$ through the instrument branch, which minimally disturbs the quantities on the original circuit to be measured.

To measure the current i_C , the instrument would be used as an ammeter with minimal resistance in series with R_C .

For power system installations, it is often necessary to scale down the measurement quantities by several orders of magnitude with some kind of *transducer*. The standard tool for this purpose is an *instrument transformer*, where the turns ratio is designed to deliver a safe and practical current and voltage to the instrument. Current transformers are known as CTs, and voltage or potential transformers as VTs or PTs (these terms are synonymous). The basic idea is illustrated for a sample application in Figure 16.3.

As an alternative to a transformer, voltage can be scaled down for measurement purposes with a *voltage divider*. In a voltage divider, the physically measured voltage is a fraction of the voltage of interest, determined by the ratio of impedances on the measured circuit branch and the one of interest. This ratio can be known very precisely. In Figure 16.4,

$$V_m = \frac{Z_2}{Z_1 + Z_2} V_s$$

Capacitors are more practical than resistors for this purpose because they dissipate much less energy. Given that it is still necessary to make contact with high-voltage equipment somewhere, transducers (and their installation labor) can dominate the cost of sensing equipment in the electric grid. In the context of high-precision measurements, instrument transformers may also dominate the measurement error.

16.2.3 Reporting Measurements

After stepping voltages or currents down to a reasonable scale for measurement, another crucial step is analog-to-digital (A-to-D) conversion. This step encodes the analog value observed by the instrument at some instant into some number of bits, depending on the desired precision, that can

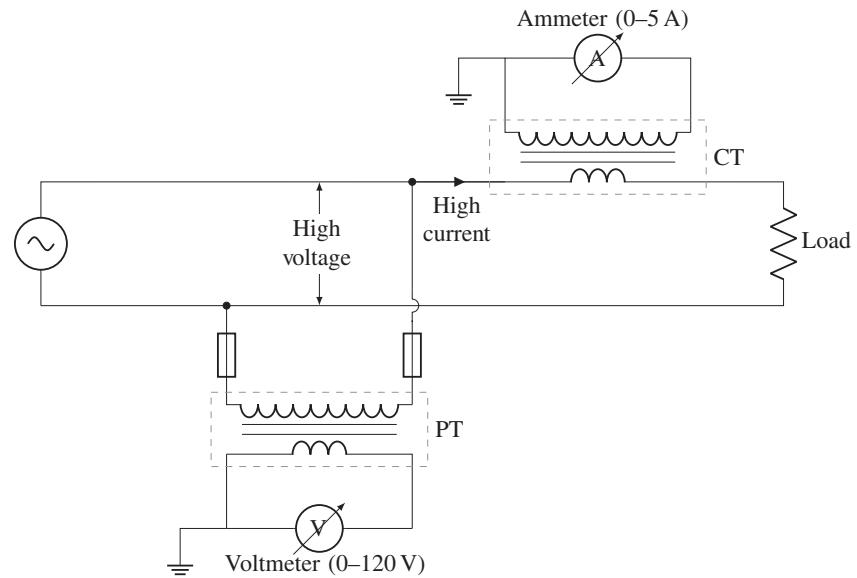


Figure 16.3 Potential and current transformers.

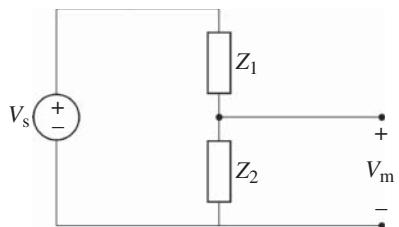


Figure 16.4 Voltage divider.

then be stored, assigned a time stamp, and used in further computation. Turning physical measurement values into a series of reported data frames involves summarizing or compressing information into chunks that are manageable and useful for the purpose at hand. Inevitably, this data compression will be lossy, and it can also produce artifacts. Thus, it is important to match expectations of a measurement to the assumptions made in producing it. We will return to this point in the context of PMUs.

The most common type of a.c. measurement data reported is a root-mean-square magnitude value (see Section 3.1.3), averaged over some time interval. The shortest meaningful interval for an rms value is half a cycle, but most digital instruments report such values less frequently (say, every few seconds).

Sometimes the detailed waveform is reported as *point-on-wave* (POW) measurements. This will reflect any deviations from the ideal sinusoidal waveform and yield information about the *harmonic content* (Section 5.3), depending on the number of samples per cycle. The two areas of grid operation where POW measurements have historically been used are power quality analysis and protection. In both cases, the measurements are strictly local, whether for diagnostic purposes (power quality) or taking direct action (protective devices).

Until recently, it was not feasible or practical to store long time series of POW data, much less communicate them in real time. For example, a power quality expert might view a waveform with a portable instrument on location, and save a few cycles' worth of data in a screenshot

or summary. A *digital fault recorder* (DFR) is a device that automatically stores detailed local waveform measurements for a limited duration (say, seconds to minutes) when triggered by a disturbance. This strategy is known as *report by exception*. DFRs often accompany protective devices at the transmission level, and this type of data is used for general diagnostics as well as postmortem analysis in the event of an outage.

As information and communication technologies evolve, and accordingly notions of what constitutes large and cumbersome data volumes, it is becoming increasingly feasible and practical to record synchronized, *continuous point-on-wave* (CPOW) measurements on a monitoring basis. The ability to directly observe departures from the assumed sinusoidal waveform under both routine and nonroutine conditions may prove useful for a growing number of applications.¹⁰

16.2.4 PMUs

Phasor measurement units (PMUs) are an important tool for observing and analyzing power systems. The idea is to represent voltage or current *phasors* in the same format that we conceptualize as electric power engineers. As introduced in Section 3.5, a phasor compresses the time dimension and represents a quantity that is assumed to vary sinusoidally as a snapshot mapped into the complex plane. This formulation preserves two aspects of the waveform: root-mean-square magnitude, and phase angle relative to some zero reference. Different sinusoidal quantities of the same frequency (as we would ideally expect from all voltages and currents in an a.c. network) can then be easily compared against each other and analyzed graphically. The voltage phasor plays an especially important role because the voltage phase angle difference is associated with real power transfer between locations in an a.c. network (see Sections 9.3.4 and 12.4.1).

To identify phasor differences between two locations, it is necessary to synchronize measurements to a common time reference, which is why PMU data are also called *synchrophasors*. Note that by itself, a phase angle from a single point on a circuit has no physical meaning, since one can call time $t = 0$ whenever. Access to GPS satellite-based time signals in the 1980s made it possible to share a precise reference time stamp, and thus compare the timing of measurements at arbitrary locations.¹¹ To get a sense of the required resolution, the time interval corresponding to 1° of angle at 60 Hz is 46.3 μs .¹² Of course, due to other factors contributing toward the error budget, a much more accurate time stamp is needed to yield data with, say, a 1° resolution.

According to industry standards (repeatedly revised since the early 2000s), most PMUs report rms magnitudes, phase angles, local frequency (the rate of change of phase angle), and rate of change of frequency (ROCOF). Typical reporting rates are one or two data frames per cycle, or once every other cycle. These quantities are calculated from analog measurements taken at a much higher internal sampling rate. For further data economy, three-phase measurements are often compressed into a single *positive-sequence* value that characterizes the behavior of three balanced phases (see Section 4.2).

¹⁰ See A. Silverstein and J. Follum, *High-Resolution, Time-Synchronized Grid Monitoring Devices* (North American Synchrophasor Initiative and Pacific Northwest National Laboratory, 2020). Available: https://www.naspi.org/sites/default/files/reference:documents/pnnl_29770_naspi_hires_synch_grid_devices_20200320.pdf (accessed April 2023).

¹¹ Prior to high-precision global timekeeping services, it would have been necessary to use expensive atomic clocks at each location. Today, several alternative networked timing references exist, and chip-scale atomic clocks are also becoming practical. GPS remains an easily accessible common standard across many industries and consumer products.

¹² The GPS time signal itself is accurate to tens of nanoseconds, but its interpretation is limited by the quality of the receiver. In the case of PMUs, measurement error in both magnitude and angle may be dominated by the instrument transformers used to access grid-level voltages and currents.

PMU data have provided unique empirical insights into both steady-state operation across wide areas of interconnected a.c. systems, and into disturbances.¹³ It is now routine for system operators to monitor the voltage phase angle separation across important transmission links, especially those that are stability constrained (see Section 7.3.2).

PMU measurements also revealed the presence of sub-synchronous (i.e., slower than 50 or 60 Hz) oscillations across a.c. systems, especially in response to sudden disturbance events. Figure 16.5 illustrates local frequency as computed by the rate of change of the local phase angle at four different locations in Texas, some hundreds of miles apart, following the sudden loss of a generator.¹⁴ The rough overall shape of the frequency excursion is that conceptualized in the context of load frequency control (Section 11.1), where we assume that a.c. frequency is the same everywhere. However, within that shape are also oscillations due to generator dynamics as described by the swing equation (see Section 13.4.4). This phenomenon had never been directly observable prior to synchronized phasor measurements.

A fundamental problem with PMU measurements is that they are based on the assumption of a sinusoidal signal. As mentioned in the derivation of phasor notation in Section 3.5.2 and shown in Eq. (3.37), the argument of the sinusoidal function can be split into a rotating phasor $e^{j\omega t}$ and a stationary phasor $e^{j\phi}$, where the rotating phasor (with information about frequency) is simply discarded while the stationary phasor provides the snapshot that is conventionally called the “phasor.”

As we have seen, a.c. frequency in real power systems is not, in fact, constant. Even when the signal appears roughly as a sine wave, we are not dealing with a true mathematical sinusoid, which is perfectly periodic to infinity. Therefore, neither frequency nor phase angle are strictly defined. For example, “frequency” of an approximately periodic signal might reasonably mean the inverse of time between zero crossings, the rate of change of phase angle, the maximum energy bin of a Fourier transform, or the best fit auto-correlation in time. For realistic waveforms, different algorithms for computing frequency may produce slightly different results.

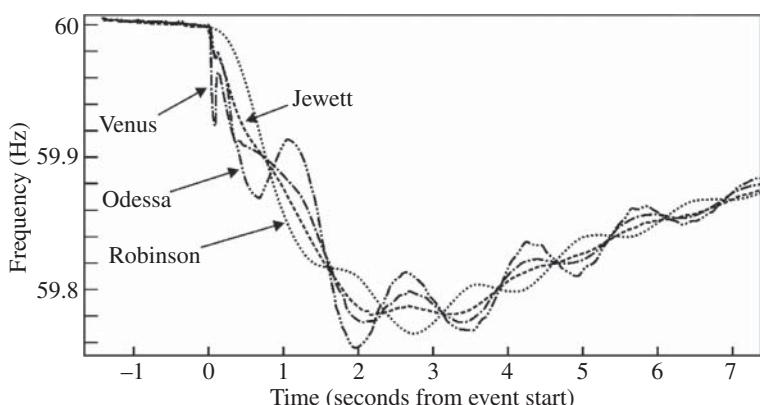


Figure 16.5 Oscillations in local a.c. frequency due to generators swinging against each other, rendered observable with early PMUs. Source: From Faulk & Murphy (1994).

¹³ Live frequency and phase angle data can be seen at <https://fnetpublic.utk.edu/> (accessed April 2023), courtesy of the University of Tennessee, Knoxville.

¹⁴ The original graph, reproduced in Figure 16.5, was presented in a stunning report by D. Faulk and R. Murphy, “Comanche Peak Unit No.2 100% Load Rejection Test - Underfrequency and Systems Phasors Measured Across TU Electric System,” *47th Annual Conference for Protective Relay Engineers* (College Station, TX: Texas A&M University, 1994).

When the waveform is smooth and frequency changes only slowly, it is easy to agree on an approximate representation, as we have done implicitly throughout this book and power engineers have done for well over a century. However, it is important to recognize that there is no such thing as a “true” value of frequency or phase; the best we can do is to agree upon a consistent process by which to make a measurement.

Even more important, in situations where the waveform is significantly disturbed, calculated quantities can be misleading if taken literally. For example, Figure 16.6 illustrates waveforms on three phases as recorded during a fault event.¹⁵ One would be hard pressed to identify a “correct” phase angle or frequency in this picture. The model of a sinusoid to which we are trying to make the data conform is simply not apt here.

One strategy for resolving the above difficulties is to reconsider phasor estimation as a curve fitting problem. In fitting a signal—that is, a series of rapid analog measurements over some time interval—to the model

$$x(t) = X_{\max} \cos(\omega t + \phi)$$

(where x may refer to voltage or current), we are asking: which numerical values of the three parameters X_{\max} , ω , and ϕ would yield the closest prediction of the observed data?¹⁶ One obvious advantage of this approach is that we acknowledge upfront our ignorance of the frequency ω during the observation interval. Another advantage is that the curve-fitting process (e.g., a least-squares optimization) will produce another metric called the *goodness-of-fit*, which represents the error quantity to be minimized by the best choice of parameters. This goodness-of-fit can serve as an indication or flag as to whether our sinusoidal model was apt for the interval in question. The model used in curve-fitting could also include a fourth, explicit parameter C for the rate of change

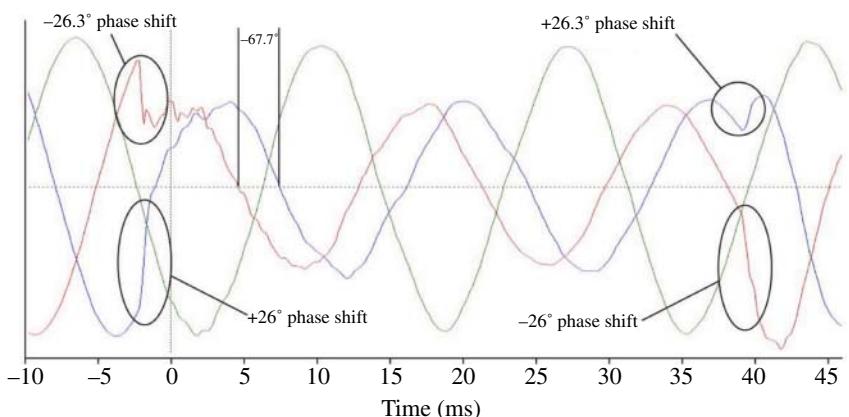


Figure 16.6 Recorded waveforms during a transmission fault. Source: NERC (2017)/with permission of North American Electric Reliability Corporation.

¹⁵ The 2016 Blue Cut Fire incident, see also Section 14.4.3 about inverter response. Source: North American Electric Reliability Corporation (NERC), *1,200 MW Fault Induced Solar Photovoltaic Resource Interruption Disturbance Report, Southern California 8/16/2016 Event*, June 2017.

¹⁶ The curve-fitting approach has been proposed by Harold Kirkham and collaborators. For example, see H. Kirkham (2016), “The Measurand: The Problem of Frequency,” *IEEE International Instrumentation and Measurement Technology Conference Proceedings (I2MTC)*, May 23–26, 2016, Taipei, Taiwan. DOI: 10.1109/I2MTC.2016.7520347.

of frequency,

$$x(t) = X_{\max} \cos\left(\frac{1}{2}Ct^2 + \omega t + \phi\right)$$

In sum, a variety of algorithms can be applied to the phasor measurement problem, best understood as an attempt to fit observations to an abstract model that can never be a perfect description of reality.

16.2.5 Actionable Intelligence and Automation

Over the course of recent decades, electric grid operations and control centers have cautiously adopted digital information technology to manage and convey growing amounts of data. Transmission systems have taken a much higher priority in this respect than distribution systems, for good reasons. First, much more is at stake: any given node at the transmission level represents more megawatts, more money, and more customers subject to an outage. The possibility of large-scale blackouts resulting from some oversight is ample reason to invest in any technology that can support reliable operations. Second, distribution systems were historically intended for one-way power flow, and assumed to be much simpler in nature. The design idea was to size distribution infrastructure to safely withstand worst-case conditions (i.e., peak loads and faults), and there would be no need to monitor them closely. The proliferation of distributed generation (Section 15.2) is radically changing this situation, although many distribution utilities may still find it challenging to make a business case for higher-resolution sensing and control.

At the highest level of operational oversight for a given geographical and jurisdictional area, a grid operator or balancing authority is responsible for the overall integrity of the system, including market transactions to be realized within that physical system. Arguably, some of the most sophisticated information and computer systems in the electric power industry relate to markets. A state-of-the-art control room at this level is shown in Figure 16.7.¹⁷ The upper row of numbers includes time of day, area control error (ACE, see Section 11.1.5), frequency, and load for the area. The various screens below capture an impressive volume of information including network connectivity, equipment operational status, and power flows at specific locations, along with weather, generation, and load forecasts.

Still, this carefully curated information represents a minuscule fraction of what it would take to fully describe the operating state of the entire power system, much less all the factors that go into predicting its future state. In part, this is because much of the detailed technical information (such as circuit breaker status or power flow on smaller lines) is handled on a smaller geographic scale, by each utility that owns transmission lines and is responsible for serving customers within its territory. Yet at each operational level, it must be considered not only what information is available, but what is *actionable*. As elaborated in Section 16.3.3, more data is not always better when it comes to making decisions in real time, since attention has a cost.

Due to the high degree of complexity and the high stakes involved, the electric power industry as a whole has been quite cautious about automating decision processes and actions. Ideally, computational tools can support human operators by sourcing relevant information and succinctly presenting it for decision-making to human operators. Such tools range from *expert systems* to more recently emerging *artificial intelligence* (AI) and *machine learning* (ML) techniques, initially for research and offline applications.¹⁸

¹⁷ Based on image owned by the CAISO and licensed with permission from the California ISO. Any statements, conclusions, summaries or other commentaries expressed herein do not reflect the opinions or endorsement of the California ISO.

¹⁸ M. Bariya, “Applications of time synchronized measurements in the electric grid (Order No. 28717978),” 2021. Available: <https://www2.eecs.berkeley.edu/Pubs/TechRpts/2021/EECS-2021-196.pdf> (accessed April 2024).



Figure 16.7 California Independent System Operator (CAISO) control room. Source: California Independent System Operator.

Constraints for many utilities and grid operators to pursue such developments include a lack of sufficiently large and accessible empirical data sets, as well as time for technical staff to experiment with novel techniques. While power generation technology tends to be fairly standardized and thus lends itself better to automated control processes, transmission and distribution systems are highly idiosyncratic and dependent on local knowledge.

Overall, by the standards of many other industries in the 21st century, electric power systems lag far behind in their utilization of state-of-the-art digital platforms and data. As the cost of information decreases in relation to the cost of energy, it increasingly makes economic sense to carefully survey the grid and intervene in its operation through an increasing number of data points and control nodes, deferring to computers where necessary to handle the information volume or achieve the required reaction speed. For example, in the context of distributed generation, the sheer volume of information and control opportunities associated with so many generators and diverse behaviors would seem to require some form of automation if resources are to be actively monitored, coordinated, and their contributions optimized. Nevertheless, reasons to proceed with caution include not only economics but overall complexity, operational risk, and human factors.

16.3 Human Factors

16.3.1 Operators and Engineers

This section characterizes some of the important differences between an academic and a practical view of power systems.¹⁹ Specifically, it distinguishes between operations and engineering as two types of human expertise and activity that are essential to the grid's performance. The following is

¹⁹ Parts of this analysis have been previously published in A. von Meier, "Occupational Cultures as a Challenge to Technological Innovation," *IEEE Transactions on Engineering Management* 46(1), 101–114, February 1999.

a summary of findings from research conducted by the author that set out to explain differences of opinion about the introduction of automation technology in distribution systems by examining how different people conceptualize the power system as a whole. Although technology options have changed substantially over the past decades, human cognition has not, and neither have the core roles of operators and engineers.²⁰ Concerns about technologies such as novel computer interfaces and remote control in the 1990s may well resonate with today's concerns about the introduction of artificial intelligence (AI) and machine learning (ML) applications in power systems.

A majority of readers, it is assumed, will be more familiar with engineering as an activity and modeling framework. In the electric power industry, "engineering" encompasses a great variety of specific job tasks. Engineers make design drawings, calculate specifications, select components, evaluate performance, and analyze problems. Their work has an important idealistic aspect, finding innovative solutions and always striving to improve things. Some utility engineers are directly engaged with physical hardware (e.g., overseeing its installation); others work with abstract models of the power system (e.g., power flow analysis) or on its indirect aspects (e.g., instrumentation or computer systems). Those engineers whose work is more remote from the field and of a more academic nature best match the archetype of the present description.

Operators of technical systems, be they power plants, airplanes, or air traffic control, must keep the system working in real time. In a thermal generation plant, for example, operators have to assure that steam flows at a multitude of temperatures and pressures between boiler, turbine, and condenser remain coordinated and in balance with electrical load on the generator unit throughout a range of conditions. In electric power transmission and distribution, operators monitor and direct ongoing reconfigurations of their system of interconnected power lines and components from switching stations and in the field. Unlike engineering, where the object is to optimize performance, the goal in operations is to maintain the system in a state of equilibrium or homeostasis in the face of external disturbances, steering clear of calamities. An operating success is to operate without incident. Depending on the particular system, maintaining such an equilibrium may be more or less difficult, and the consequences of failure more or less severe.

Three types of challenges are generally characteristic of the operations job: external influences, clustering of events, and uncertainties in real-time system status. Especially in the case of power distribution systems, a large part of the hardware is physically accessible and vulnerable to all kinds of disturbances, whether they are automobiles crashing into poles or foxes electrocuting themselves on substation circuit breakers. Events such as heavy storms or extreme loading conditions entail cascading effects in the system and require a large number of switching, diagnostic, and repair operations to be coordinated and carried out under time pressure. At the same time, system parameters such as loading status for certain areas or even hardware capabilities are often not exactly known in real time. Operators are quite accustomed to working in this sort of situation, and the way they view and imagine the electric grid, along with their values and criteria for system performance, can be seen as specific adaptations to these challenges.

16.3.2 Cognitive Representations of Power Systems

In the engineering framework, "the system" is considered as a composite of individual pieces, since these are the units that are readily described, understood, and manipulated. The functioning of the system as a whole is understood as the result of the functioning of these individual components:

²⁰ "Operators" and "engineers" are understood here as archetypes, not rigid or exclusive groupings. The boundaries of these cultural groups may not coincide exactly with departmental boundaries or formal occupational titles held by individuals; in fact, some individuals may serve in dual roles and share each archetypal perspective to various degrees.

should the system not work, the obvious first step is to ask which component failed. Engineering is therefore analytic, not only in the colloquial sense of investigating a complex thing, but analytic in the very literal sense of “taking apart,” or treating something in terms of its separate elements.

Like any analytic process, engineering requires modeling, or representing the actual physical system in abstracted and appropriately simplified terms that can be understood and manipulated. Abstraction and simplification also requires that the system elements be somehow idealized: each element is represented with its most important characteristics, and only those characteristics, intact. An engineering model will thus tend to consider system components in terms of their specified design parameters and functions. Each component is assumed to work as it should; components with identical specifications are assumed to be identical. Similarly, the relationships among components are idealized in that only the most important or obvious paths of interaction (generally the *intended* paths) are incorporated into the model. The parameters describing components and their interactions are thought of as essentially time-invariant, and invariant with respect to conditions not explicitly linked to these parameters.

The behavior of the system is thus abstracted and described in terms of formal rules, derived from the idealized component characteristics and interactions. These rules, combined with information about initial conditions, make the system predictable: from the engineering point of view, it should be possible in principle to know exactly what the system will do at any point in the future, as long as all rules and boundary conditions are known with sufficient accuracy. These rules also imply a well-understood causality: it is assumed that things happen if and only if there is a reason for them to happen. Of course, engineers know that there are random and unpredictable events, but in order to design and build a technical system, it is essential to be able to understand and interpret its behavior in terms of cause-and-effect relationships. Chains of causality are generally hierarchical, like if-then decision-making systems. Stochasticity is relegated to well-delimited problem areas that are approached with probabilistic analysis.

In summary, then, the classic engineering representation of a technical system can be characterized as *abstract*, *analytic*, *formal*, and *deterministic*. By contrast, the operator representation of a technical system can be typified as *physical*, *holistic*, *empirical*, and *fuzzy*. This representation is instrumental to operators in two important ways: it lends itself to maintaining an acute situational awareness, and it supports the use of intuitive reasoning.

Because operations involve much more immediate contact with the hardware, system components are imagined as the real, physical artifacts in the way that they are perceived through all the senses. For example, a particular overhead distribution switch has a certain dimension, offers a certain resistance to being moved, makes a certain noise, and shakes the pole in a certain way as it closes. Even when looking at abstract depictions of these artifacts on a drawing or a computer screen, operators “see” the real thing behind the picture. With all its physical properties considered, each artifact has much more of a unique individuality than its abstract representation would suggest: one transformer may overheat more than another of the same rating, or one relay may trip slightly faster than another at the same setting. Thus, components that look the same on a drawing are not necessarily identical to an operator.

To be sure, operators must also work with abstract representations. For distribution operators, for example, this means primarily circuit maps and schematic diagrams for switching. However, the abstractions they find useful and transparent may differ from those preferred by engineers. While good maps for engineers are those that do a thorough job of depicting selected objects and their formal relationships, the most useful maps for experienced operators are those that most effectively recall their physical image of the territory.

Another aspect of operators' cognitive representation is that they conceptualize the system more as a whole than in terms of individual pieces. Rather than considering the interactions among components as individual pathways that can be isolated, the classic operator model is of one entire network phenomenon. Every action taken somewhere must be assumed to have repercussions elsewhere in the system, even if no direct interaction mechanism is known or understood. This is consistent with operators' experience, where they are often confronted with unanticipated or unexplained interactions throughout the system.

Rather than using formal rules to predict system behavior, operators rely primarily on a phenomenological understanding of the system, based on empirical observation. The underlying notion is that no number of rules and amount of data can completely and reliably capture the actual complexity of the system. Therefore, though one can make some good guesses, one cannot really know what will happen until one has seen it happen. No component can be expected to function according to its specifications until it has been proven to do so, and the effect of any modification has to be demonstrated to be believed. While engineers would tend to assume that something will work according to the rules, even if it did not in the past, operators expect that it will work the way it did in the past, even if analysis suggests otherwise. Many arguments between engineers and operators can be traced to this fundamental difference in reasoning.

Finally, the operator representation is one that expects uncertainty rather than deterministic outcomes. Whether due to the physical characteristics of the system, insufficiency of available data, lack of a complete understanding of the system, or simply external influences, uncertainty or "fuzziness" is taken to be inevitable and, to some degree, omnipresent. Ambiguity, rather than being subject to confinement, is seen to pervade the entire system, and operators suspect the unsuspected at every turn. Thus, distribution operators have described their system as a "live, undulating organism" that must somehow be managed.

This physical, holistic, empirical, and fuzzy view of the system is adaptive to the challenge of operating the system in real time in that it allows one to quickly condense a vast spectrum of information, including gaps and data pieces with different degrees of uncertainty, into an overall impression or *gestalt* that can be consulted with relative confidence to guide immediate action.

Finally, operators tend to draw on intuitive reasoning, especially when data are insufficient but action is required nonetheless. Although there are manuals specifying operating procedures, many situations occur that could not have been foreseen in detail and courses of action recommended. To deal with the problem at hand, analytic tools may not be able to provide answers quickly enough. Worse yet, information on the books may be found to be untrustworthy under the circumstances; for example, if recent data appear to contradict what was thought to be known about the system. In order to come to a quick decision, the operator's main recourse then is to recall past experience with similar situations. How did the system behave then? Were people surprised? How did the particular equipment respond? Based on such experience, an operator will have an intuitive "feel" for the likelihood of success of a given procedure.

This experience-based approach is intuitive not because it is irrational, but because it is nonalgorithmic. An operator might have difficulty articulating all the factors taken into consideration for such a decision, and how, precisely, they were mentally weighed and combined. He or she might not be able to cite the reasons for feeling that something will work or not work. Nonetheless, the decision makes use of factual data and logical cause-effect relationships, as they have been empirically observed.

The use of intuitive processes is so deeply embodied in the culture of operations that they are often chosen over analytic approaches by preference rather than necessity. Obviously, both methods can fail; the question is about relative degrees of confidence. While engineers may frown on

operator justifications that seem based on intractable, obscure logic or even superstition, operators delight in offering accounts of situations where their intuition turned out to be more accurate than an engineer's prediction. In fact, both approaches are adaptive to the work contexts of their proponents, and while both have a certain validity, either approach may turn out to yield better results in a given situation. The important point here is that substantive differences in cognitive representations and reasoning modes underlie what may appear to be trivial conflicts or petty competition between cultural groups. These differences will also have specific implications for the evaluation of system design, operating strategies, and technological innovations.

16.3.3 Operational Criteria

The most important general properties of technical systems, or goals and criteria for evaluating their performance, can be summarized as *efficiency*, *reliability*, and *safety*. These goals tend to be shared widely and across subcultures throughout an organization managing such a system. However, individuals or groups may hold different interpretations of what these general goals mean in practice and how they can best be realized. Accordingly, they will also have different expectations regarding the promise of particular innovations.

When there are trade-offs among safety, reliability, and efficiency, cultural groups may also emphasize different concerns, not only because they have different priorities, but because they have different perceptions of how well various criteria are currently being met. In the academic engineering context, it is often assumed that certain standards of safety and reliability have already been achieved, and the creative emphasis is placed on improving efficiency. In the case of power systems, safety and reliability are problems that were academically solved a long time ago, whereas new approaches to increase efficiency offer a continuing intellectual challenge.

The efficiency criterion thus takes a special place in engineering. Efficiency here can be taken in its specific energy-related sense as the ratio of energy or kilowatt-hour output to energy input, or in a more general sense as the relationship of output, production, or benefit to input, materials, effort, or cost. Efficiency is often a direct performance criterion in that its numerator and denominator are crucial variables of interest that appear on the company's "bottom line" (e.g., electric generation and revenues). Even where efficiency measures something more limited or obscure (e.g., how many man-hours are required for service restoration), a more efficient system will generally be able to deliver higher performance at less cost while meeting the applicable constraints. Conversely, low efficiency indicates waste or the presence of imperfections that motivate further engineering. A more efficient system will also be considered more elegant: beyond all its practical implications, efficiency is an aesthetic criterion.

In addition, there is a set of indirect or supporting criteria that, according to the cognitive framework of engineering, advance efficiency as well as safety and reliability. While these criteria can be taken as qualitative standards for the system as a whole, they also apply in evaluating technological innovations and judging their promise. One such criterion is *speed*. It is an indirect criterion because it does not represent an actual need or an immediate, measurable benefit. However, the speed of various processes offers some indication of how well the system is theoretically able or likely to succeed in being efficient. Generally, a system that operates faster will involve less waste. For example, restoring service more quickly means less waste of time, waste of man-hours, and waste of potential revenues. Responding and adapting to changes faster can also mean higher efficiency in terms of improved service quality or saved energy. Given the choice between a slow- and a fast-operating device, all else being equal, most engineers would tend to prefer the faster one.

Similarly, *precision* is generally considered desirable in engineering culture. Actually, the desired criterion is *accuracy*: not only should information be given with a high level of detail, but it should be known to be correct to that level. Accurate measurements of system variables allow for less waste and thus support efficiency; they also further safe and reliable operation. However, the accuracy of a given piece of data is not known *a priori* and is subject to external disturbances, while its degree of precision is obvious and inherent in design (e.g., the number of significant figures on a digital readout). Precision can be chosen; accuracy cannot. Although precision does not guarantee accuracy, it at least provides for the possibility of accuracy and is therefore often taken in its place (and sometimes confused). Given the choice between a less and a more precise indicator of system parameters or variables, most engineers would prefer the more precise one.

More fundamentally, *information* in and of itself is desirable. Generally, the more information is available, the better the system can be optimized, and information can in many ways advance safety and reliability as well. In the event that there are excess data that cannot be used for the purpose at hand, the cost to an engineer of discarding these data is typically very low: skipping a page, scrolling down a screen, or ignoring a number is no trouble in most engineering work. In selecting hardware or software applications, all else being equal, most engineers would prefer those offering more information.

Finally, the ability to control a system and its parts is another indication of how successfully the system can be engineered, managed, and optimized. This is because any variable that can be manipulated can also, in principle, be improved. As with information, in the engineering context, there is hardly such a thing as too much control. If the ability to control something is available but not needed, the engineer can simply ignore it. Most engineers would prefer to design systems and choose components that are controllable to a higher degree.

This set of criteria suggests a general direction for technological innovations that would be considered desirable and expected to perform well. Specifically, from the viewpoint of engineering, innovations that offer increased operational speed, precision, information, and control appear as likely candidates to further the overall system goals of efficiency, reliability, and safety. While such expectations are logical given the representational framework of engineering, the perspective of operations yields quite a different picture.

Of the three general system criteria—safety, reliability, and efficiency—safety takes a special priority in operations, while efficiency is less of a tangible concern. From the point of view of managing the system in real time, efficiency is an artifact of analysis and evaluation, a number tagged on after the fact, having little to do with reality as it presents itself here and now. Although it may indicate operating success, efficiency more directly measures the performance of engineers. Most operators would agree that having an efficient system is nice, as long as it does not interfere with their job.

Safety, on the other hand, takes on a profoundly tangible meaning for operators because the consequences of errors face them with such immediacy. In power systems, any single operation, performed at the wrong time, has the potential to cause customers to lose power. Immediately, telephones will ring, voices on the other end will shout and complain, and the control room may even fill with anxious supervisors. Because of the interdependence of power system components, the consequences may occur on a much larger scale than the initial error. Aside from causing power outages, incorrect switching operations can damage utility and customer equipment.

But even more serious is the risk of injury or death. Any steam generation plant, for example, harbors the intrinsic hazard of water vapor at up to 1000°F and several hundred atmospheres of pressure. It is difficult to stand next to a roaring primary steam valve without sensing that the smallest leak could be instantly fatal. In transmission and distribution, the risk of electrocution looms for utility crews as well as others who might be accidentally in contact with equipment

(e.g., people in a car under downed power lines). The one action T&D operators dread most in the course of switching operations is to energize a piece of equipment that is still touching a person. Like operators of other technical systems, they carry a personal burden of responsibility for injuries or fatalities during their shift that goes far beyond their legal or procedural accountability. The difference between an intellectual recognition and the direct experience of the hazards cannot be overemphasized: hearing an accident described is not the same as watching one's buddy die in a flash of sparks a few feet away, with the smell of burning flesh. This immediate awareness of the life-taking potential of electricity is omnipresent among operators of power systems. Implicitly or explicitly, it enters any judgment call they make, whether about day-to-day procedures or about implementing new technology.

Their acute perception of safety colors operators' interpretation of other system goals and helps define their criteria for good system design and performance. The set of criteria—speed, precision, information, and control—that, from the engineering perspective, support not only efficiency but also safety and reliability may be seen by operators as less important or even counterproductive. Instead, operators value a different set of criteria that specifically support their ability to operate the system safely.

Speed, generally advantageous in engineering, is more problematic in operations because one is working in real time. Speed is desired by operators in the context of obtaining information. They may also wish for their actions to be executable quickly, so as to gain flexibility in coordinating operations. However, a system of fast-responding components and quickly executed operating procedures, where effects of actions propagate faster and perhaps farther, also introduces problems: it will tend to be less tractable for the operator, provide less time to observe and evaluate events and think in between actions, and allow problems to become more severe before they can be corrected. Power systems are inherently fast in that electric effects and disturbances propagate at the speed of light, making cascades of trips and blackouts practically instantaneous. Any delays or buffering of such effects work toward the operator's benefit. Thus, from the perspective of operations, *stability* is generally more desirable than speed. Operators would prefer a system that predictably remains in its state, or moves from equilibrium only slowly, allowing for a greater chance to intervene and bring it back into balance.

Information can also be problematic in the context of operations. To be sure, there are many examples of information that operators say they wish they had, or had more of. But more is not always better. Because one is gathering information and acting upon it in real time, the cost of discarding irrelevant information is not negligible. Deciding which data are important and which are not costs time and mental effort. By increasing *cognitive load*, too much data can distract from what is critical and may interfere with operators' acute situational awareness. Operators often give examples of information overload: many computer screens that must be scanned for a few relevant messages, or many pages of printout reporting on a single outage event. Generally, instead of greater quantity of information, operators desire *transparency*, meaning that the available information is readily interpreted and placed into context. It is more important for them to maintain an overview of the behavior of the whole system than to have detailed knowledge about its components: in terms of maintaining situational awareness, it is preferable to lack a data point than to be confused about the big picture even for an instant. If more information has the potential to create confusion, then for operators it is bad.

Suppose, for example, that a computer screen is to display real-time measurements from throughout the system to operators in a control room. One option might maximize information delivery, say, providing constantly updated figures from 100 sensor nodes. In a situation where all 100 data points are equally likely to be relevant, where it is important that no detail be missed,

and where the data need not be processed and acted upon with great time pressure, this may be most desirable. By contrast, suppose that much of this information is irrelevant to decisions that must be made very quickly. Here it may be appropriate to reduce the amount of information in the interest of transparency, for example, by limiting the number of points reported, or by displaying only those that changed recently. The idea is that data should quickly and correctly characterize the situation behind the numbers, even at the expense of breadth or depth.

Similarly, more precision is not always better for operators. While engineers can make use of numbers with many significant digits, the last decimal places are probably not useful for guiding operating decisions. In fact, operator culture fosters a certain skepticism of any information, especially quantitative. This skepticism is consistent with their keen awareness of the possibility of foul-ups like mistaking one number for another, misplacing a decimal point, or trusting a faulty instrument, and the grave potential consequences. Therefore, operators' primary and explicit concern about any given numerical datum is whether it basically tells the true story, not how well it tells it. Moreover, precision can be distracting or even misleading, suggesting greater accuracy than is in fact given. Thus, in operations, *veracity* of information is emphasized over precision. Rather than trusting a precise piece of information and running the chance of it being wrong, operators would generally prefer to base decisions on a reliable confidence interval, even if it is wide.

The difference between precision and veracity is that precision offers a narrow explicit margin of error, while veracity offers confidence that the value in question truly lies within that margin, and that the value really represents that which it is assumed to represent. Suppose one measurement is taken with a crude and foolproof device, while another is displayed by a sophisticated monitoring network. The latter offers more precision, but the skeptic may wonder: Is it possible that the instrument is connected to the wrong node? Could the display be off by an order of magnitude? Or might it be telling me yesterday's value instead of today's? And if so, would there be any discernible warning? If someone's life depended on our correct estimate of a quantity, we would prefer the former, more reliable measurement, even though it is less precise.

Finally, more control is not always better. Of course, there may be variables over which operators wish they had more control. But the crucial difference is that in engineering, control always represents an *option*, whereas in operations there may be an associated responsibility to exercise this control: the ability to control a variable can create the expectation that it *should* be controlled, and produce pressure to act. Operators tend to be wary of such pressure, primarily because it runs counter to a basic attitude of conservatism fostered by their culture: When in doubt, don't touch anything. Their reluctance to take action unless it is clearly necessary arises from the awareness that any operation represents a potential error, with potentially severe consequences. An interventionist approach that may allow greater optimization and fine-tuning thus inherently threatens what operators see as their mission, namely, to avoid calamities.

In pragmatic terms, more controlling options may mean that operators have more to do and keep in mind, and thereby increase stress levels. Alternatively, they may not have time to exercise the control at all, in which case their performance will be implicitly devalued by the increased expectation. Because time and attention are limited resources in operations, and because of the potential for error associated with any action, the option not to control can be more desirable than the ability to control. This option is provided by a system's *robustness*, or its tendency to stay in a viable equilibrium by itself.

Suppose that a technological innovation gives operators the ability to control some system parameter (voltage, for example) within a narrower band, closer to the desired norm. The potential

downside of this innovation has to do with the following question: What happens if, for whatever reason, the control option is not exercised? For example, is the parameter liable to drift farther outside of the normal range if it is not actively controlled? In doing so, does it pose a safety hazard? Does the new technology raise expectations for system performance, leading to disappointment if control actions are not taken in the manner envisioned by system designers? Will pressure to exercise control options create extra work for operators? By contrast, robustness and stability characterize a situation where things will be fine without active intervention.

Another example of efficiency versus robustness in the power-distribution context is the handling of load peaks in view of limited equipment capacity. An “efficient” approach might call for transferring loads in real time among various pieces of equipment so as to achieve the most even distribution and avoid overloading any one piece. This approach both maximizes asset utilization (and may even help avoid capacity upgrades to accommodate demand growth) and minimizes the inefficiency due to lines losses (since collective I^2R losses increase with uneven allocation of current among lines). Its success hinges, however, on constant vigilance and intervention. By contrast, a “robust” approach would emphasize strength and simplicity: the idea is simply to have enough extra capacity built into the equipment so that overloading is not an issue, and loads need not be tracked so carefully.

In summary, then, the system qualities that are most important for operators are stability, transparency, veracity, and robustness, which support them in their task of keeping the system in homeostasis. Not coincidentally, these criteria are generally associated with older technologies, designed and built in an era where operability by humans was an absolute requirement. In power systems, stability and robustness were provided largely by oversized equipment and redundancy of components, while transparency and veracity were furnished through simple mechanical and analog instrumentation and controls. From the viewpoint of increasing the efficiency of such systems in today’s world, process innovations guided by engineering criteria may be desirable indeed. From the operations perspective, however, such innovations may be expected to adversely affect performance reliability and especially safety. Thus, when steps are proposed toward more refined and sophisticated system operation, operators may identify potential backlash effects, in which opportunities for system improvement also introduce new vulnerabilities.

16.3.4 Implications for Technological Innovation

The cognitive representations used by engineers and operators, respectively, give rise to different ideas about what system modifications may be desirable, and divergent expectations for the performance of innovations. If one imagines a technical system in terms of an abstraction in which interactions among components are governed deterministically by formal and tractable rules, then (i) these formal relationships suggest ways of modifying individual system parameters so as to alter system performance in a predictable fashion according to desired criteria, and (ii) it is credible that such modifications will succeed according to *a priori* analysis of their impacts on the system. From this point of view, innovative technologies hold much positive promise and little risk.

On the other hand, if one imagines the system as an animated entity with uncertainties that can never be completely isolated and whose behavior can be only approximately understood through close familiarity, then (i) modifications are inherently less attractive because they may compromise the tractability and predictability of the system, and (ii) any innovation must be suspected of having unanticipated and possibly adverse consequences. From this point of view, efforts to modernize or automate may imply an attempt to squeeze the system into a conceptual mold it does not fit—treating an animal like a machine—and thus harbor the potential for disaster.

This is not to suggest that modernization efforts are ill-advised or necessarily conflict-ridden. Indeed, there are many examples of operators welcoming if not actively campaigning for improved and more efficient information and control technology. The preceding observations suggest that when concerns about technological innovation do arise, they may be based on legitimate professional considerations.

Consider the example of SCADA at the distribution level. Traditionally, distribution operators sitting in the control room at the distribution switching center relied on their field crews as the main source—and sometimes the only source—of information about system status, whether switches (open or closed), loads (current through a given line or transformer), voltage levels, or the operating status of various other equipment (circuit breakers, capacitors, voltage regulators, etc.). The operators' lifeline to information was the telephone or radio through which his "eyes and ears" in the field communicate. By the same method, operating orders (often written out in hardcopy beforehand) were verified, or modified orders communicated if necessary. The introduction of SCADA implied a transition from operating through field personnel to directly accessing the system via a computer terminal in the control room. The advantages are obvious: fewer man-hours are needed to execute a given procedure; things can be done much faster; the computer affords a clear, central overview—in short, the entire operation, still largely based on 19th- and early 20th-century technology, finally comes into the electronic age.

Nevertheless, the implementation of SCADA in the utility industry was not entirely unproblematic. While many distribution operators were quick to point out the advantages of SCADA and indeed found it difficult to imagine how anyone ever worked without it, some also offered thoughtful critiques. The main areas of concern relate to safety (is the computer correct in reporting an open switch?), physical surveillance (sacrificed chances to discover any developing problems early on in site visits), time pressure (not having time to think while waiting for field crews to execute orders), loss of redundancy (not necessarily having a second person reviewing steps), and the loss of situational awareness (such as that afforded by audible communication among operators, as opposed to silent interaction with computer terminals). Gaining full confidence took many years of empirical observation, habituation, and debugging.

As a result of such concerns, even when new information and control technology is successfully installed, operators may be reluctant to make full use of the available capabilities. This hesitation must be expected especially if a new technology pertains to a more automated operation that might threaten to make the system less transparent, verifiable, and robust to operators.

A related, more recent example of technology innovation in the area of sensing involves PMUs (Section 16.2.4). Synchrophasor data may be available in control rooms, but remain underutilized if they are not recognized as a clear and trustworthy source of actionable information that helps solve immediate problems. Despite their strategic role to increase situational awareness, new and unaccustomed types of data can increase cognitive load and thus be experienced as a distraction in the short run.

Finally, some challenges to the integration of inverter-based resources must be understood in this light. Quite aside from what advanced power electronic controls are capable of and how reliably they perform, there is a question of how predictable they and their effects on the system appear to operators. Much of this book describes approximations and heuristics that have made it possible for humans to manage a complex system in practice that cannot be fully accounted for in theory. Technology that is governed by different physical laws than rotating machines, along with digital controls that operate on time scales inaccessible to the human mind, threatens to make some of these old heuristics obsolete. From the standpoint of an operator whose job it is

to understand the system, its dynamics and its safety margins, this epistemological challenge is a profound peril.

Arguably, computing technology (including artificial intelligence and machine learning) could be applied and accumulate a trustworthy track record to effectively address some of these issues. It might be possible after all to delegate the comprehension of a complex system to a machine—at least in the context of offline modeling and analysis. For real-time operations, such developments will merit considerable caution. As has historically been the case, successful innovation will require ample consultation across occupational and cultural groups, to account for both engineering and operating concerns throughout the design process.

16.4 Strategic Perspectives

In parallel with technological innovation, electric grids have faced or will face two big transitions in the 21st century: economic restructuring, and climate mitigation. Clearly, these are both broad and complex topics. Far from a thorough analysis, this section merely aims to suggest how some of the technical properties of power systems discussed throughout this book inform or constrain available choices in navigating these transitions.

16.4.1 Decarbonization

The electric grid is at the center of efforts to mitigate climate change by reducing carbon emissions. Studies of transition scenarios for decarbonizing the energy sector have shown that along with energy efficiency and substitution of carbon-neutral fuels, electrification of energy end uses is critical for meeting climate goals.²¹ Basically, this is because using renewable resources to generate electricity—all the complexities and storage needs notwithstanding—is still easier than making carbon-neutral chemical fuels to replace fossil fuels on a large scale. Therefore, even with energy efficiency improvements to meet existing electric end uses with fewer kilowatt-hours (Section 6.4.4), significant overall load growth to include newly electrified end uses (especially transportation, but also heating and cooking) is widely anticipated. At the same time, shifting away from fossil generation means new coordination challenges in space and time to match resources with demand.

The associated technical evolution and its challenges can be understood in four broad categories: grid dynamics, supply–demand balancing, localization, and transmission infrastructure. Underpinnings of each of these issues have been addressed throughout this book.

The term *dynamics*, as contrasted with steady-state operation, refers to phenomena that are not described within the framework where a.c. frequency is the same everywhere and waveforms are smooth and regular. Dynamics in the legacy grid are dominated by rotating synchronous generators and their swing equation (Section 13.4), where instabilities can arise due to sudden step changes (Section 13.4.4) and vulnerability is created by large power flows (Section 7.3.2). The collective behavior of large numbers of machines interacting across a transmission network is very difficult to model, but empirical observation over many decades has provided heuristics or rules of thumb by which operators can maintain system stability. For example, it may be known

²¹ For example, Jim Williams, Andrew Debenedictis, Rebecca Ghanadan, Amber Mahone, Jack Moore, William R. Morrow, III, Sneller Price, and Margaret S. Torn, “The Technology Path to Deep Greenhouse Gas Emissions Cuts by 2050: The Pivotal Role of Electricity,” *Science* 335(6064), 53–59, 2011. DOI: 10.1126/science.1208365.

empirically that staying within some limits of power injections and flows will keep oscillations (as seen in Figure 13.6 or 16.5) within tolerable levels.

The introduction of inverter-based solar and wind generation changes the physics of the system such that old rules will no longer apply. Consequently, previously observed minor oscillations could become less well-damped and grow into serious threats, or new instabilities could be introduced by unintended interactions among inverter control loops. Very fast and precise control of inverters (Section 14.4) can, in principle, mitigate problems and even improve these dynamics—but not until the many interactions across the network are better understood and we know what, exactly, we should be telling inverters to do.

A more obvious and intuitive problem associated with growing contributions from variable renewable energy sources is matching supply and demand (Chapter 11) at all times of day, every day of the year. Solar and wind power (Sections 15.1.3 and 15.1.4) have some predictable patterns that can be planned for, as well as stochastic volatility that must be compensated for in near real time. Energy storage (Section 15.3) is a crucial solution, where different technologies are suited to address variations on different time scales (and the seasonal imbalance is likely the biggest challenge). It is important to remember, though, that the need for energy storage is negotiable to some extent when traded off against complementary solutions that include demand response (Section 6.4.4), broader regional transmission interconnection, and simply overbuilding solar or wind generating capacity. Planning tools (Section 16.1.4) and institutional arrangements to support such comprehensive, system-wide least-cost optimization are still evolving.

In spatial terms, the transition to renewable resources—specifically, the explosive growth of solar photovoltaics (PV)—also implies a transition to more local, distributed generation (Section 15.2) that presents both challenges and opportunities. Distribution systems (Section 7.1) may require capacity and protection (Section 7.5) upgrades to simply accommodate new resources along with increased loads. They may also benefit from more refined visibility (Section 16.2) and control (Section 16.2.5) to coordinate resources locally, with new options to create microgrids (Section 15.4) or ad-hoc power islands (Section 7.1.6) for resilience.

Finally, the integration of solar and wind power on a large scale hinges on expanding transmission infrastructure between load centers and remote areas with the most abundant and concentrated resources (including offshore wind). This entails technical consideration of loading limits (Section 7.3) with stability and security analysis (Section 13.3), as well as policy challenges beyond the scope of this text.

16.4.2 Markets

A transition from vertically integrated, regulated regional monopolies toward various forms of competitive marketplaces for electricity has been underway in many countries roughly since the turn of the millennium. Implementations vary by region and jurisdiction, and markets along with regulatory design continue to evolve. As of this writing, the U.S. map resembles a patchwork quilt of different rules and arrangements, with some areas essentially unchanged and others on the cutting edge of experimentation.

Electricity markets may include any aspect of production, delivery, and consumption, although the primary focus has been on wholesale energy. Academic and political discourse has seen spirited debates about the comparative merits and failings of government regulation versus free, competitive markets as applied to different transactions involving electricity. Most implementations are in fact some hybrid, whether by embedding economic incentives within a regulated system, or by

governmental definition of rules and boundary conditions for competitive markets. One point on which experts agree is that the consequences to society are important and potentially far-reaching.

Because electricity is such a central and indispensable part of our culture, the topic goes far beyond kilowatt-hours and dollars. Our choice of means and institutions for dealing with electricity touches upon fundamental and contentious questions about how society and business ought to be organized. A key set of questions concerns the extent to which competitive markets can produce socially optimal outcomes, what constitutes market failures, and when and how government intervention is required. At issue is an entire *weltanschauung* about appropriate roles of public and private sector, depending in large part on one's degree of faith in the "invisible hand" of the marketplace. The context of a transition to a renewable, decarbonized grid puts an even sharper point and greater urgency on these considerations.

Far from a thorough coverage of electricity market and regulation topics, this section is intended only to highlight some of the intrinsic technical challenges presented by electric grids with respect to market function.²² Physical characteristics and constraints in its production, distribution, and consumption make electricity quite different from other goods in the economics textbook. The simplistic notion of a commodities spot market, where the intersection of demand and supply curves yields an equilibrium price and quantity at which the market clears (and thereby maximizes collective benefit to participants), is grossly insufficient for understanding or managing electric power systems. Consequently, electricity market design has seen many modifications and added layers to try to account for the complexity and to correct shortcomings.

Some of the basic requirements of an ideal competitive market are that demand and supply can vary freely with respect to price, and that units of a good offered by different suppliers are interchangeable. Electricity fails to meet these criteria for several different reasons.

First, demand for electricity is not generally very well characterized as a response to price signals (one might say the *price elasticity of demand* is low, or poorly defined), in part simply because electricity is so indispensable to our economy and our everyday lives. People and businesses have been organized around being able to consume electricity at any time they choose, and some are willing to pay considerable amounts of money to maintain this freedom and not to have their service interrupted (Section 13.2.2). Crucially, the *flexibility* to tailor one's use of electricity to circumstances in real time has a value distinct from the *quantity* of energy consumed in kilowatt-hours.

Indeed, the entire planning and design philosophy of power systems and regulated utilities in the 20th century presumed that electricity should be available to anyone in essentially arbitrary amounts around the clock, at a known and fixed price deemed reasonable by public regulators. Electric load was defined as the independent variable whose satisfaction became the central objective of technical, organizational, and regulatory efforts throughout the power industry. Legacy grids included no way for electric utilities to remotely shut off individual customers or throttle their power usage, other than crude rotating outage blocks at the level of entire distribution circuits.

Consumers experienced rate changes only on the time scale of years, from one utility rate case to another, rather than a "price signal" according to the much more rapid fluctuations of marginal generation cost. In many markets, consumers remain shielded from short-term price volatility and therefore receive no direct incentive to respond to power shortages, making demand almost perfectly inelastic with respect to varying wholesale prices.

Market solutions introduced to varying degrees in different jurisdictions include *real-time pricing* and related incentive mechanisms to encourage electric *demand response* (Section 6.4.4). These can

²² An excellent reference for further reading is Daniel S. Kirschen and G. Strbac, *Fundamentals of Power System Economics* (John Wiley & Sons, 2nd edition, 2018).

be considered extensions or refinements of well-established mechanisms such as *time-of-use rates* for predetermined on- and off-peak periods, or *interruptible tariffs*. Nevertheless, routine implementation of real-time demand responsiveness on a large scale is a considerable innovation. It means turning upside down a century-old design philosophy and requires addressing a broad set of factors ranging from customer behavior, education, and economic preferences to control hardware, information management, and communications protocols—factors that had no reason to be considered in electric power systems until quite recently. Thus, while it seems plausible in theory to imagine aggregate electric demand as a well-behaved function of price, such a transition in practice requires new ways of thinking as well as novel use of technology.

The supply side harbors equally if not more serious challenges. A competitive market according to the economics textbook requires free entry and exit of firms in response to scarce or excessive supply. Under *direct access*, an arrangement aimed at fostering competition in the generation sector while preserving a T&D monopoly, any generator can in principle avail themselves of the extant infrastructure by injecting electric power to be consumed by a purchaser elsewhere.

Yet in practice, market entry and exit are impeded by capital intensity, transaction costs, permitting, and construction schedules for larger projects. Historically, the capital-intensive nature of electric generation along with economies of scale and a measure of risk have made this industry uninviting for smaller businesses. Interconnection arrangements, too, may be cumbersome if not prohibitive for small producers, with transaction costs including fees, bureaucratic procedures, risk of uncertain timing, and special equipment to interface with the utility on its terms.

To foster market participation by small-scale distributed energy resources (Section 15.2), FERC Order 2222 encouraged the formation of aggregator entities as a vehicle for them to participate in regional wholesale power markets.²³ The concept of *transactive energy systems* extends the vision of market engagement down to the local level and the device scale, with direct negotiation among diverse resources including generation, load, and storage, facilitated by modern communication and digital control technology.²⁴ The idea is that by reconciling power and energy needs, availability, and costs via automated processes close to prosumers, strain on the grid can be relieved and overall efficiency improved. At the same time, it remains illegal anywhere in the United States to run independent lines from a supplier to a consumer—say, a wire from my rooftop PV system to my next-door neighbor—without becoming a regulated public utility.²⁵

Aside from transaction cost, the time scale on which traditional power plants are planned, built, and licensed makes it all but impossible for new capacity to appear in response to acute shortages. Existing facilities are limited both in terms of their rated capacity in megawatts²⁶ and their ramp rate in megawatts per minute, depending on the type of plant (Section 15.1). Consequently, the supply curve becomes less elastic as the number of megawatts required at a particular moment approach the system's limit. In practice, when demand threatens to exhaust available supply, system operators are left to scramble for any sources that might be able to eke out another few megawatts of

23 Order No. 2222 was issued by the U.S. Federal Energy Regulatory Commission in 2020, with updates in 2021. See <https://www.ferc.gov/ferc-order-no-2222-explainer-facilitating-participation-electricity-markets-distributed-energy> (accessed October 2023).

24 See, for example, GridWise Architecture Council, *Transactive Energy Research, Development and Deployment Roadmap* (2018). PNNL-26778, <https://www.osti.gov/servlets/purl/1573922> (accessed October 2023).

25 This restriction has been motivated by safety as well as the intention to avoid the inefficiency of duplicating infrastructure, though it is also arguably the greatest barrier to a truly competitive electricity market. One can imagine that the proliferation of distributed generation technologies will ultimately challenge this status quo; if so, it would amount to nothing short of a radical reinvention of “the grid” as we know it.

26 It is often possible for an electric generator to exceed its nameplate rating at the expense of equipment life span. Nevertheless, every generation unit ultimately encounters a nonnegotiable physical limit on the amount of power it can deliver.

additional supply or load reduction, no matter the cost. These desperate phone calls are also termed *out-of-market settlements*.

Not only is electricity supply constrained, but so is the grid's ability to transmit it to the desired location. With increased utilization of transmission capacity and long-distance sharing of generation resources, transmission congestion has become as important as generation in limiting the available supply for many areas. At the same time, upgrading transmission capacity tends to be slow, expensive, and politically contentious, so that bottlenecks are not quickly relieved by building more lines (Section 16.1.4). Most important, congestion cannot be made to go away by administratively or economically "rerouting" power: as previous chapters (especially Section 2.3 and Chapter 12) have emphasized, electric power flow obeys only physical law; it is difficult enough to predict, let alone control and direct. Market approaches for addressing locational constraints include *zonal* or *nodal pricing*, which aim to reconcile supply and demand with increasingly fine spatial resolution—in effect, having a different wholesale clearing price in each geographic area.

A key concern in market design has been preventing the potential exercise of *market power* by sellers. At issue is whether market participants are *price takers* as in the textbook, or whether by their individual actions they have the power to influence the market clearing price and thus manipulate outcomes in their favor—*gaming* the market. This may not require a large market share. For example, because line flows are sensitive to the amount of power injected at specific locations, generation capacity can be used strategically by scheming market participants not only to relieve but to deliberately create transmission congestion.²⁷ The unique problem lies both in the utter dependence of the entire system on limited transmission assets, and the subtlety of predicting power flow. The effect of generation at a given node on congestion at a certain link is sometimes obvious from geography, but the details are carefully guarded by system operators to prevent abuse. The resultant paradox is that key information about the system must be withheld rather than openly shared in the interest of competitive market function, which is precisely the opposite of what we would expect from a textbook situation.

Another way to understand the profound challenges in the design of electricity markets is to observe that megawatt-hours produced at different locations and different times are not generally interchangeable. In fact, not even megawatt-hours injected at the same time and location are interchangeable, because their cost and value are context-dependent in other ways. Specifically, the cost of supplying and the value of consuming electricity varies as a function both of the recent history and the predicted future of supply and consumption by a given market participant. For example, a generator with a limited thermal ramp rate expecting to sell power in the morning must stay hot and spinning at night, even if the electricity cannot be sold at a profitable price. Likewise, the cost or inconvenience of reducing electric demand at a particular hour could hinge on whether a building was precooled or a battery charged with advance notice. For all these reasons, electricity is not a *commodity* (like wheat, pork bellies, or petroleum) as found in the economics textbook. This means there is no guarantee that one supplier's product can be substituted for that of another, violating a central tenet of competitive markets.²⁸

In view of these complications, and especially in the context of growing contributions from renewable resources, it may well become appropriate to quantify buying and selling actions in terms other than units of energy. Throughout the era of fossil fuels, it made perfect sense to conceptualize energy as a stock quantity like tons of coal, barrels of oil, or cubic feet of natural gas. To the extent

²⁷ Transmission congestion may raise regional wholesale prices, of which a company or financial interest group can take advantage if they also own generation facilities located in that region.

²⁸ For a theoretical treatment of this and related issues, see Leigh Tesfatsion, *A New Swing-Contract Design for Wholesale Power Markets* (Hoboken, NJ: John Wiley & Sons, 2021).

that production costs were dominated by fuel and other variable operational costs, it was logical to trade in units of tangible product like kilowatt-hours, which is conserved as a physical quantity and readily measured as a transaction. But the cost of solar and wind farms is almost entirely capital cost, independent of how much energy is actually injected into the grid. At the same time, there is likely a growing need for resources to provide various forms of balancing services or flexibility to the grid.

Ancillary services include various types of generation reserve, distinguished by the expected response time, duration and magnitude of contribution; examples include spinning reserve, load following, and frequency regulation (Section 11.1). Reactive power and voltage support are another category of ancillary services (Section 16.1.3). The common theme is that resources are recruited and paid for their actions in response to real-time conditions on the grid, helping steer toward a state of balance, while quantities of net energy transfer are of lesser concern.

It stands to reason that in largely decarbonized systems with little or no intrinsic energy cost, such balancing or flexibility services will become increasingly central to electric power markets. Systems depending on variable renewable resources will also require more services specifically to manage uncertainty in the face of imperfect weather and resource forecasts on various time scales. The notion of insurance services may be more apt for this purpose than traditional energy contracts.²⁹ The complexity of trading these grid services in the face of the subtle yet firm physical constraints of a.c. power flow appears unmatched by any other market or set of goods.³⁰

Beyond making the system work in real time, policy makers also look to energy and power markets to create appropriate incentives for investment, and to spur the long-term evolution deemed in the best interest of society. It is safe to say that the intersection of strategic decarbonization and market design poses a set of challenges that reasonable people disagree how best to navigate, depending in part on the envisioned timeline and degree of perceived urgency.

This brings us, finally, to a very basic problem pervading the energy sector: the *external costs* of many of its components, primarily the environmental impacts of various generation sources, that fail to be included in the apparent costs of production and consumption. Externalities associated with different technologies span a wide range of issues that are often difficult to monetize or weigh comparatively—from local air pollution, toxicity, water quality, land use, aesthetics, mining, and habitat destruction for endangered species to national security and the global and intergenerational specter of climate change.

Economists broadly agree that the presence of externalities warrants some form of regulatory intervention, whether framed as financial incentives or explicit requirements, aiming to *internalize* these costs in the market as much as possible and allow participants to make well-informed decisions. The “correct” price would then promote activities known to produce social benefit while discouraging those with hidden costs to society. Arguably, when considering a long-term strategic transition of the entire energy sector, investments that would be too slow in coming from the private sector have to be made or facilitated by government. Examples of interventions that have received much discussion in recent years include a carbon tax, portfolio standards for renewable generation resources, and publicly funded infrastructure projects such as transmission lines. Suffice it to say that experts differ on the type and extent of methods favored.

29 See L. Tesfatsion, “Locational Marginal Pricing: When and Why Not?,” Economics Working Paper No. 23003 (Ames, Iowa: Iowa State University, 2023). <https://www2.econ.iastate.edu/tesfatsi/LMPWhenAndWhyNot.LTesfatsion.pdf> (accessed September 2023).

30 For example, financial markets involve many complicated derivative products and sophisticated trading strategies—but the crucial difference is that these abstract products, unlike electricity, are not subject to physical laws that might suddenly invalidate every transaction across the entire network.

The list of problems outlined in this section does not rule out the successful application of market mechanisms to electricity in principle. Rather, it emphasizes that electric power systems pose serious, inherent challenges to the design of markets and policies that will actually produce the intended results in both the short run and long run.

Ultimately, a central problem for power systems has always been equity. Electric grids have historically served as instruments for the reallocation of significant amounts of wealth across society—not in a revolutionary way, but very much in line with cultural norms, like the graduated income tax. In the 1930s, federally subsidized rural electrification brought villages and farms onto the grid that never could have been connected economically, as revenues from these accounts clearly would not pay for the infrastructure investment. In fact, many of those rural areas might have been far more cost-effectively supplied with local sources such as wind power. But the electric grid was designed to reach out even to customers who were expensive to serve, while splitting the price tag among everyone. In this way, urban customers continue to subsidize rural customers, and large commercial and industrial customers, who are much less expensive to serve on a per-kWh basis, continue to subsidize smaller customers under regulated utility rates. A key impetus for restructuring was the desire by such larger customers to procure less expensive electricity on their own through open market access, prompting considerable argument as to their obligation to share the cost burden of past utility investments. The debate about how to organize the electric industry has always involved questions about what is fair, who pays how much, and how society should decide.

The historical dimension brings to light just how much the complex technical system we call “the electric grid” is a social artifact as much as a fascinating incarnation of physics and engineering. We can view the electrification of industrial society not only as a logical result of technical and economic driving forces, but as an embodiment of ideas and values. The idea of connectedness in and of itself represented the sophistication and unity of an advanced, civilized society: being “on the grid” meant not only receiving electrons, but being part of progress in the modern world. By the same token, living “off the grid” has implied independence, self-sufficiency, and renouncing the downsides of industrialization.

Historically, the idea of equal access to the grid has represented a fundamental sense of social equality and togetherness, for better or for worse, with access to electricity considered an entitlement of all citizens. Today, the electric grid is at the center of monumental efforts to address climate change, once again—or still—an indispensable piece for articulating visions of the future. As power systems evolve with new technologies and new organization, they will continue to embody social values, explicit or implicit. The better and the more widely these technical systems in their complexity are understood, the greater the opportunity for people to guide this evolution with awareness and conscious choice.

Appendix A

Symbols, Units, Abbreviations, and Acronyms

A	ampere	unit of current
A		transmission line parameter
A		symbol for area
a		symbol for an operator that performs 120° rotation
A.C., a.c.		alternating current
ACE		area control error
AGC		automatic generation control
Ah	ampere-hour	unit of stored charge
AM		amplitude modulation
amp	ampere	unit of current
ASD		adjustable speed drive
B		transmission line parameter
B		symbol for magnetic flux density
B, b		symbol for susceptance
B _f		symbol for frequency bias constant
BJT		bipolar junction transistor
BTM		behind the meter
BTU		British thermal unit, unit of energy
BWR		boiling water reactor
C	coulomb	unit of charge
C		transmission line parameter
C		symbol for cost
C, c		symbol for capacitance
c		symbol for speed of light
CAES		compressed-air energy storage
CAISO		California Independent System Operator
cal	calorie	unit of energy
CHP		combined heat and power
CPOW		continuous point-on-wave
CSC		current-source converter

CSP		concentrating solar power
CT		current transformer
CVR		conservation voltage reduction
D		transmission line parameter
<i>D</i>		symbol for damping constant in swing equation
<i>D, d</i>		symbol for distance
D.C., d.c.		direct current
DFIG		doubly fed induction generator
DFR		digital fault recorder
DG		distributed generation
DQZ, dqz		direct, quadrature, zero components
<i>e</i>		natural logarithm base, 2.71828...
<i>E, e</i>		symbol for voltage
<i>E</i>		symbol for electric field
E		symbol for energy
ELF		extremely low-frequency fields
emf, <i>emf</i>		electromotive force
EMF		electromagnetic fields
erg		unit of energy
EUE		expected unserved energy
F		symbol for force
F	farad	unit of capacitance
<i>f</i>		symbol for frequency
FACTS		flexible a.c. transmission systems
FERC		Federal Energy Regulatory Commission
FET		field effect transistor
FIDVR		fault induced delayed voltage recovery
FM		frequency modulation
G	gauss	unit of magnetic flux density
<i>G, g</i>		symbol for conductance
<i>g</i>		symbol for gravitational acceleration
GFCI		ground-fault circuit interrupter
GMD		geometric mean distance
GMR		geometric mean radius
H	henry	unit of inductance
<i>H</i>		symbol for magnetic field strength
<i>H</i>		symbol for generator inertia constant
<i>h</i>		symbol for height
hp	horsepower	unit of mechanical power
HVDC		high-voltage direct current
Hz	hertz	unit of frequency
<i>i</i>		imaginary number, $\sqrt{-1}$

<i>i</i>		summation index
<i>I, i</i>		symbol for current
IEEE		Institute of Electrical and Electronics Engineers
IGBT		insulated gate bipolar transistor
ISO		independent system operator
ITIC		Information Technology Industry Council
<i>j</i>		imaginary number, $\sqrt{-1}$
J	joule	unit of energy
<i>J</i>		symbol for rotational inertia
K	Kelvin	unit of absolute temperature
<i>k</i>		summation index
KCL		Kirchhoff's current law
KVL		Kirchhoff's voltage law
kV	kilovolt	unit of potential difference, 1000 V
kVA	kilovolt-ampere	unit of apparent power
kW	kilowatt	unit of power, 1000 W
kWh	kilowatt-hour	unit of energy, 1000 watt-hours
L	liter	unit of volume
<i>L, l</i>		symbol for inductance
<i>L</i>		symbol for Lagrangian
<i>l</i>		symbol for length
LCD		liquid crystal display
LDC		load duration curve
LED		light-emitting diode
LOLE		loss-of-load expectation
LOLP		loss-of-load probability
LTC		load tap changer
LWR		light water reactor
<i>M</i>		symbol for inertia constant in swing equation
<i>m</i>		symbol for mass
mmf, <i>mmf</i>		magnetomotive force
mmf		magnetomotive force
MOSFET		metal oxide semiconductor field effect transistor
MPP		maximum power point
MRI		magnetic resonance imaging
MVA	megavolt ampere	unit of apparent power
MVAR	megavolt-ampere reactive	unit of reactive power
MW	megawatt	unit of power, one million watts
MWh	megawatt-hour	unit of energy, one million watt-hours
N	newton	unit of force
<i>n</i>		symbol for rotational rate
<i>N, n</i>		symbol for number of turns in a coil

N.C., n.c.		normally closed
NERC		North American Electric Reliability Corporation
N.O., n.o.		normally open
OCB		oil circuit breaker
OLTC		on-load tap changer
OPF		optimal power flow
<i>P</i>		symbol for real power
PCB		polychlorinated biphenyls
p.f., <i>p.f.</i>		power factor
PLL		phase-lock loop
PMU		phasor measurement unit
POW		point-on-wave
PT		potential transformer
p.u.		per unit
PV		photovoltaic
PWM		pulse-width modulation
PWR		pressurized water reactor
Q		symbol for reactive power
<i>Q, q</i>		symbol for charge
\mathcal{R}		symbol for reluctance
<i>R, r</i>		symbol for resistance
<i>R, r</i>		symbol for radius
<i>R</i>		symbol for regulation constant
rad	radian	unit of angle
REC		recloser
rms		root mean squared
ROCOF		rate of change of frequency
rpm		revolutions per minute
rps		revolutions per second
<i>S</i>		symbol for complex power
SCADA		supervisory control and data acquisition
SE		state estimation
SF ₆		sulfur hexafluoride
SIL		surge impedance loading
SMES		superconducting magnetic energy storage
SOC		state of charge
SVC		static VAR compensator
T	tesla	unit of magnetic flux density
<i>T</i>		symbol for temperature
<i>T, t</i>		symbol for time
T&D		transmission and distribution
TCL		thermostatically controlled load
THD		total harmonic distortion

TOU		time-of-use
UPS		uninterruptible power supply
V	volt	unit of voltage
V, v		symbol for voltage
v		symbol for velocity
VA	volt-ampere	unit of apparent power
VAR	volt-ampere reactive	unit of reactive power
VR		voltage regulation
VSC		voltage source converter
VSD		variable speed drive
VT		voltage transformer
V2G		vehicle-to-grid
W	watt	unit of real (active) power
Wb	weber	unit of magnetic flux
X, x		symbol for reactance
Y	wye	type of three-phase connection
Y, y		symbol for admittance
y		symbol for admittance per unit length
Z, z		symbol for impedance
z		symbol for impedance per unit length
ZIP		constant impedance, current and power loads
α	alpha	symbol for an operator that performs 120° rotation
α	alpha	symbol for real part of propagation constant
α	alpha	symbol for thyristor firing angle
β	beta	symbol for area frequency response
β	beta	symbol for imaginary part of propagation constant
γ	gamma	symbol for propagation constant
Δ	delta	type of three-phase connection
Δ	delta	symbol for difference
δ	delta	symbol for angle
δ	delta	symbol for skin depth
ϵ, ε	epsilon	symbol for electrical permittivity
η	eta	symbol for efficiency
θ	theta	symbol for angle
λ	lambda	symbol for flux linkage
λ	lambda	symbol for wavelength
λ	lambda	symbol for incremental cost in Lagrangian
μ	mu	symbol for magnetic permeability
μ	micro-	one-millionth
ν	nu	symbol for frequency of a wave
π	pi	ratio of a circle's circumference to diameter, 3.1415...
ρ	rho	symbol for resistivity

ρ	rho	symbol for density
σ	sigma	symbol for conductivity
σ	sigma	symbol for surface-charge density
σ	sigma	symbol for Stefan-Boltzmann constant
Φ, ϕ	phi	symbol for magnetic flux
ϕ	phi	symbol for phase angle
Ω	ohm	unit of resistance
ω	omega	symbol for angular frequency

Index

a

- Abstraction 1, 17, 30, 41, 42, 128, 210, 214, 258, 323, 471, 477
 Accuracy 38, 251, 265, 323, 355, 463, 471, 474, 476
 Adapter 142, 145
 Adjacency matrix 342
 Adjustable speed drive (ASD) 141, 487
 Admittance 38, 70–71, 73–77, 83, 86, 91, 92, 184, 211, 213–216, 220, 236–237, 240, 242–244, 246–247, 250, 252, 326, 340
 matrix 37–38, 323, 341–344, 350, 358–359
 Advanced metering infrastructure (AMI) 461
 Algebra, linear 38, 112, 239, 349, 354, 359, 365
 Algorithm 150, 155, 314, 316, 322, 345, 353–354, 357, 361, 362, 390, 420–422, 441, 445, 450, 457–458, 461, 466, 468, 472
 Alternating current (a.c.) 9, 12, 26, 29, 33, 55–60, 62–64, 68, 69, 73–74, 101, 117, 132, 136, 159, 178–179, 197–198, 200, 203, 212, 227, 258, 260, 267, 287–290, 321, 405, 413, 415, 419, 436, 487
 Aluminum core steel reinforced (ACSR)
 conductor 52–53, 234, 254
 Ambiguity 17, 122, 397, 422, 450, 472
 Ammeter 63, 321, 463
 Ampacity 182
 Ampere (unit) 6, 8, 21, 61, 66, 182, 218, 331, 350, 408, 487
 Ampère's law 225, 229–230, 233, 234
 Amp-hours 445, 487
 Amplitude 56, 58–59, 62–63, 65, 80, 89, 104–105, 107, 130, 198, 200, 249, 416, 417, 487
 Ancillary service(s) 155, 299, 306, 307, 458, 484

- Angle stability 183–184, 376, 378–392
 Angular frequency 58, 201, 289, 492
 Anode 406–408, 413
 Antenna 25
 Apparent power 79, 81, 82, 84, 85, 110, 131, 182–184, 219, 244, 283, 284, 333–335, 423
 Arc 406–408
 Area control error (ACE) 126, 306–311, 319, 468
 Armature
 current and load 271–274, 276–277, 279, 281, 285–286, 292, 296
 reaction 261, 264, 266–267, 270, 277, 291–293
 winding 101, 265–266, 268, 271–274, 278–279, 281, 284, 288, 291–293, 381–382
 Artificial intelligence 453, 470, 479
 Asymmetric fault current 197–200
 Attenuation constant 249
 Automatic generation control (AGC) 306, 308, 309, 311, 455, 456
 Automatic voltage regulator (AVR) 276, 282, 329, 449, 460
 Automation 361, 468–470
 Autotransformer 188, 208–210
 Availability 167, 313, 370, 426, 432, 440, 450, 458, 482
 Avalanche diode 411

b

- Bandgap 409, 433
 Baseload 312, 313, 431
 Batteries 26, 42, 142, 143, 155, 406, 437, 444–448, 450, 451, 454
 Behind-the-meter (BTM) generation 154, 441
 Betz limit 435

- Biomass 429–430
 Bipolar transistor 412
 Birds 12, 432
 Blackout(s) 125, 183, 270, 310, 311, 319, 393, 403, 457, 468, 475
 Black start 270, 413, 414, 436
 Blue loss 433
 Bottoming cycle 429
 Breakdown voltage 411
 Brownout 12, 122
 Bundled conductor(s) 176–177, 179, 232–234, 236, 237, 253, 254
 Bus(es)
 P,V and P,Q 327, 329, 331, 335, 345, 350
 slack 316, 327–332, 334–335, 345–346, 350, 353–355, 358–360, 459
 swing 328
 Bus admittance matrix 323, 341–344, 350, 359
 Busbar 270, 322
- C**
- California Independent System Operator (CAISO) 152, 469
 Capacitance
 parallel or shunt 85–86, 188, 239, 243
 series 86, 133, 188
 of transmission lines 68, 84, 86, 175, 188, 225, 234–238, 245–246, 344
 Capacitor(s) 29–30, 67–70, 76–79, 81, 85–86, 92, 93, 129, 135, 149, 166, 188–190, 206, 289, 326, 334, 363, 393, 397, 418, 419, 424, 436, 439, 448, 451, 453, 463, 478
 control of 188
 Capacity (of equipment) 182, 360, 439, 444, 477
 Capacity factor 26, 154, 432
 Carnot efficiency 428, 429
 Cathode 406–408, 413
 Rays 408
 Characteristic impedance 184, 185, 248–250, 252, 253
 Characteristic wavelength 246
 Charge
 electric 1, 4, 18–22, 36, 48, 67, 231, 234–235, 257, 409–410, 448
 static 10–11
 Charging current 245
- Circuit
 behavior and analysis of 51, 323, 326
 breakers 15, 105, 110, 122, 141, 164, 166, 169–170, 190–192, 194–196, 206, 280, 371, 385, 440, 453, 457, 460–461, 468, 470, 478
 closed 11, 48, 106, 411, 418
 elements 7, 29–31, 33–34, 37, 49, 51, 60, 73–74, 80–82, 92, 96, 135, 142–143, 214, 293–294, 326, 411, 463
 lumped 7
 magnetic 48–51, 292
 open 11, 15, 40, 42–43, 45, 47, 68, 215–216, 250, 293, 418, 433–434
 short. *See* Short circuit
 Circulating current 45, 170, 173, 189, 208, 210, 279–282, 378, 380, 381
 between generators 280
 Coaxial cable(s) 175
 Coercivity 212, 213
 Cogeneration 429, 438
 Collector
 in solar plant 432, 434
 in transistor 412, 413
 Combined cycle plants 429
 Combined heat and power (CHP) 429, 438
 Communication 149, 155, 156, 246, 302, 336, 362, 447, 451, 461, 462, 465, 478, 482
 Commutation 140, 141, 262, 415
 Complex conjugate 73, 94, 339, 341, 366, 367, 395
 Complexity 34, 39, 171, 197, 232, 296, 312, 315, 321, 336, 339, 353, 361, 453, 454, 468, 469, 472, 481, 484, 485
 Complex numbers 71–73, 91, 96
 Complex plane 72–76, 81–83, 85, 87–90, 92–94, 98, 112, 114, 216, 217, 241, 245, 249, 252, 296, 465
 Complex power 77–86, 94, 95, 110–111, 223, 294, 295, 339, 341, 353, 364, 366, 367, 394, 395, 397
 Compressed air energy storage (CAES) 447
 Computer(s)
 as loads 122
 use in power flow calculation 322, 361
 Computer Business Equipment Manufacturers Association (CBEMA) curve 123n5

- Concentrating solar power (CSP) 432
See also Solar thermal power
- Condenser (in thermal power plant) 427, 470
- Condenser, synchronous. *See* Synchronous condenser
- Conductance 9–10, 31–34, 38, 43, 50, 70, 75, 76, 96, 175, 211–213, 242, 243, 252, 341, 344, 356, 357
- Conductivity 4–5, 9, 49, 174, 178, 181, 212, 225, 408
- Conductor(s)
- bundling 176
 - diameter 231, 269
 - heating 131, 135–136, 174, 182, 269, 284, 285
- Congestion 171, 182, 483
- Conservation voltage reduction (CVR) 123–125, 149
- Contingencies 173, 307, 308, 374–376, 402
- Contingency analysis 357, 375, 376
- Continuous point-on-wave (CPOW) 465
- Control
- of inverters 480
 - of synchronous generators 381
- Cooling towers 427
- Coordination
- of protective devices. *See* Protection coordination
 - of resources 479
- Copper losses 205
- Corona 175–177, 179, 181, 232, 237, 240
- Coulomb (unit) 2, 3, 6, 68, 72, 445
- Coulombic efficiency 445
- Cross product 21, 22
- Culture(s) 150, 373, 462, 472–474, 476, 481
- Current
- alternating. *See* Alternating current
 - circulating. *See* Circulating current
 - direct. *See* Direct current
 - direction of 6, 96, 136, 245, 264, 413–414, 418
 - induced. *See* Induced current
 - inrush. *See* Inrush current
 - in power flow analysis 321, 325, 343
 - propagation speed of 7, 250
 - starting. *See* Starting current
- Current source converter (CSC) 412, 413, 415, 418
- Cycle(s) 55–56, 58, 62, 64, 66, 69, 76–79, 81, 86–88, 94, 102, 104, 123, 125, 126, 135, 137–138, 151, 161, 189, 191–193, 197–201, 213, 252, 260, 265–268, 273–274, 280, 287, 288, 330, 333, 336, 381, 384, 387, 391–392, 406, 407, 414–418, 420, 424, 427, 429, 431, 443–445, 448, 455, 464
- d**
- Damping force 382, 384, 386
- Degrees of angle 296
- Delta and wye connections
- floating 110
 - grounding of 206
 - in transformers 206–207
- Demand
- coincident and non-coincident 151
 - elasticity of 481
- Demand response 154, 155, 370, 480, 481
- Demand side management (DSM) 154
- Deregulation 459
- Derivative 65, 66, 69, 90, 247, 312, 315–317, 345–349, 351, 355–357, 384
- Partial 314, 315, 317, 345, 346, 351, 355–357
- Device 7–8, 10–11, 13–14, 29, 30, 35, 38, 42–43, 49, 50–51, 60, 63–64, 66, 67, 73, 75–76, 78, 80, 81, 85, 111, 121, 123, 127–129, 131–133, 135–137, 142, 143, 146–147, 149–150, 155, 166, 169, 187–188, 190–197, 200, 203, 205, 211, 216–217, 234, 257, 271, 280, 291, 343, 377, 390, 394, 396, 405–409, 412–416, 419, 424, 428, 434, 438, 439, 446, 448–449, 451, 455, 462, 464, 473, 476, 482
- Dielectric material 10, 68
- Differential equation 201, 246–248, 382
- Digital fault recorder (DFR) 465
- Dimension 1, 58, 84, 87, 175, 183, 195, 219, 226, 235, 240, 241, 248, 272, 300, 312, 325, 338, 343, 344, 376, 384, 388, 399, 424, 442, 465, 471, 485
- Dimmer 12, 137–139
- Diodes 30, 44, 128, 137, 408, 409, 418, 419
- Direct access 482
- Direct and quadrature components 117–119

- Direct current (d.c.)
 circuit(s) 29–53, 63, 137, 321, 324–325, 407
 circuit breaker(s) 192
 generator(s) 139, 261–262, 269–270, 406
 motors(s) 136, 140, 257, 262, 406
 transmission 163, 175, 179–180
- Dispatch 153, 155, 161, 306, 312–320, 327–330, 332, 350, 360, 361, 376, 432, 440, 443, 450, 456–458, 460
- Distortion power factor 85, 131–132, 134, 143
- Distributed generation 112, 170, 171, 363–365, 374, 437–442, 468, 469, 480
 impacts on distribution systems 439–440, 442
- Distributed line parameters 239, 246, 247
- Distribution, primary and secondary 145, 206, 207, 222
- Distribution system
 looped 169, 171–173
 network 169, 170
 radial 122, 168
- Droop control 301–304, 439, 449, 460
- Droop curve 301–305, 309, 310, 318, 319, 366, 422, 423
- Duck curve 153, 434
- Dynamic stability 338, 377, 382, 384, 385
- e**
- Earth 4, 17, 104, 144, 229, 238
See also Ground
- Economic dispatch 306, 312–320, 450, 458
- Economics 150, 153, 180, 312, 429, 432, 437, 457, 459, 469, 481–483
- Economies of scale 105, 159, 160, 432, 437, 448, 482
- Edison, Thomas 16, 55, 159, 179, 405
- Efficiency 26, 27, 101, 125, 132, 139–141, 143, 144, 154, 163, 173, 174, 181, 203, 205, 254, 265, 269, 297, 314, 406, 412, 416, 417, 424, 425, 427–430, 432, 433, 435–438, 439, 443–445, 448, 473–475, 477, 479, 482
- Electrification 159, 440, 479, 485
- Electrocution 190, 440, 474
See also Shock
- Electromagnet 21–26, 48, 53, 63, 70, 139, 140, 142, 203, 257, 264, 283, 285, 287–290, 405, 426
- Electromotive force (emf) 11, 21
 in generator 126, 183, 272
 in transformer 56, 204, 212, 295
- Electrons 2–6, 8–11, 13, 14, 19, 20, 22, 24, 25, 43, 60, 62, 68, 78, 139, 192, 232, 257, 258, 307, 406–412, 414, 417, 433, 437, 444, 445, 485
- Emitter 412, 413
- Energy 3, 5, 11, 13–15, 17–18, 24–26, 29, 43, 48, 51, 60–62, 66, 69, 74, 78–84, 105–111, 125, 139, 141–143, 150, 153–156, 160, 161, 163, 173, 174, 177, 181, 183, 189, 205, 209–213, 237, 257–258, 269–270, 276–277, 280, 288, 299–300, 302, 312–313, 327–328, 337, 371–373, 382–391, 406, 409, 413–416, 418, 420, 422–423, 425–435, 437–439, 442–449, 454, 455, 458, 460–461, 463, 466, 469, 473, 479–484
- Energy conservation 42, 64, 81, 83, 124, 209, 242, 245, 260, 261, 280, 288, 307, 382, 383, 386, 454
- Engineering 1, 5, 9, 16, 49, 58, 69, 71, 81, 89, 106, 117, 128, 150, 177, 178, 184, 197, 222, 231, 246, 265, 283, 306, 312, 361, 369, 382, 393, 400, 405, 425, 432, 436, 451, 462, 469–471, 473–477, 479, 485
- Equal-area criterion 389
- Electric Reliability Council of Texas (ERCOT) 162
- Entropy 410, 426, 428, 433
- Equilibrium
 deep and shallow 381
 position of generator 277, 376, 382, 391
 stable and unstable 278, 377–380
- Equity 485
- Euler's equation 88–91
- Excitation 140, 185, 264, 270, 287, 289, 291, 295–298
- Exciter 264, 269, 270, 272, 276, 290, 296
- Expected unserved energy (EUE). 371
- Expert systems 468
- Exponential base 89–90
- Exponential function 89
 complex 89
- Extension cord 9, 10, 12, 38, 39, 52
- External costs 484

f

- Farad 68, 86, 97, 236
 Fault
 current 86, 110, 112, 117, 122, 188, 190–193,
 196–198, 200–202, 210, 215, 291, 440,
 451
 transient 126, 193
 Fault-induced delayed voltage recovery (FIDVR)
 149
 Feeder
 lateral 165, 191
 primary 168–170
 Field(s)
 electric 5–6, 18–19, 25, 67–70, 177, 192, 206,
 234–237, 292, 406, 408, 410–412
 electromagnetic (EMF) 23
 extremely low-frequency (ELF) 23
 magnetic. *See* Magnetic field
 Field effect transistor (FET) 412
 Field lines 18–22, 257, 259, 268, 291
 Firing angle of thyristor 414
 Flat start 345, 351, 353
 Flexible a.c. transmission systems (FACTS)
 423–424, 455
 Flicker 56, 137, 140
 Flow batteries 437, 446, 447
 Fluorescent lamps 138
 Flux, electric 235
 Flux, magnetic
 density 226–227
 in generator 51, 203–204, 257–260, 272–273,
 287–288, 291–293
 leakage 49–51, 204, 210–211, 269
 linkage 51, 64, 226, 258, 259
 Flyback diode 418
 Flywheels 447–448
 Force on a charge 22
 Forced oscillations 377
 Fossil fuels 425, 429, 430, 435, 438, 443, 447,
 449, 479, 483
 Fourier analysis 127, 129
 Four-quadrant operation 423
 Frequency
 angular 58, 201, 289, 390
 choice of a.c. 55
 control or regulation 125, 126, 162, 181,
 299–311, 406, 456, 458, 466, 484

definition and ambiguity of 23–24, 422

stability 376

tolerance 125–126, 306–307

Frequency bias constant B 308, 310

Frequency response characteristic β 304

Friction 1, 2, 4, 8, 10, 13, 141, 142, 205, 262, 263,
 286, 382, 384, 385, 426, 448

Fuel cells 415, 427, 437–439, 446, 448

Full-wave rectifier 407

Fuse(s) 190, 191, 194–198, 202, 440

g

- Galvanic isolation 209
 Galvanometer 462, 463
 Gas turbines 312, 429, 438, 453
 Gauss (unit) 21, 235
 Gauss-Seidel method 354
 Gauss's Law 234, 235, 291
 Generator(s)
 capacity of 160
 conversion of energy in 257, 406
 cooling 269
 damage 283–284, 291
 induction. *See* Induction generator
 rating. *See* Ratings
 synchronous. *See* Synchronous generator
 windings 101, 104, 269, 283, 291, 292, 295,
 296
 Geometric mean distance (GMD) 231–234, 236,
 237
 Geometric mean radius (GMR) 176, 233, 234,
 236, 237, 253, 254
 Geothermal power 430
 Gold-plating assets 372
 Governor 271, 272, 280, 299–304, 306, 310, 318,
 456
 Gradient 5, 10, 18, 139, 235, 312
 Grid, design and evolution of 442, 479
 See also Power system
 Grid-following inverter 415, 420, 449, 450
 Grid-forming inverter 449
 Ground
 conductor 61, 105, 174–175, 190, 234–235,
 237–238
 in outlet 144–145
 Ground-fault circuit interrupter (GFCI) 192
 Ground return 101, 105, 237

h

- Half-wave rectifier 407
 Harmonic(s) 85, 86, 126–133, 137, 141, 143, 212, 383, 387, 414, 416–417, 419, 424, 455, 464
 Harmonic distortion 79, 85, 121, 128, 130, 134, 416
 Heat dissipation 14, 174, 176, 187, 205, 232
See also Power dissipation
 Heat engine 139, 428, 429, 432
 Heating, resistive 13–16, 60, 131, 137, 139, 154, 161, 182, 269
See also Losses, resistive
 Helix 87, 88
 Henry (unit) 65
 Hertz (unit) 58, 267, 301, 305, 308
 High-pass filter 68n18
 High-voltage direct current (HVDC) 179–182, 415, 420, 424, 455
 Horsepower (hp) 139–141, 297
 Hydrogen
 for energy storage 437, 448
 for generator cooling 269
 Hydropower 161, 426, 444
 Hyperbolic sine and cosine 250, 253

i

- Ideal current source 43–45
 Ideal transformer 210, 211, 213–215, 218, 223, 291, 294
 Ideal voltage source 42–45, 95, 106
 Imaginary axis 72, 81, 87
 Imaginary number i or j 71
 Impedance
 characteristic. *See* Characteristic impedance matrix 342
 Impulse 1, 68, 122, 249, 426
 Incandescent lamp 14, 27, 56, 85, 136, 137, 148, 154, 460
 Incremental cost 314–318, 320
 Independent system operator (ISO) 152, 153, 162, 457, 469
 Induced current 22, 64, 66, 139, 257–261, 288
 Inductance
 as function of flux linkage 226, 229–233
 mutual 51, 116, 175, 178
 of transmission lines 175

Induction

- generator 139–140, 257, 260, 285–289, 436
 motor 85, 139–140, 149, 154, 285
 Inductor 30, 63–70, 74, 76–79, 81, 92, 93, 198, 200, 276, 326, 414, 416–418, 424

Inertia

- constant 300, 385, 390
 moment of 300
 synthetic 390, 420
 thermal 137, 155, 306

- Information 29, 34, 37–38, 59, 69, 72, 80, 82, 86, 87, 89, 106, 112–113, 122, 136, 138, 142, 155, 182, 210, 220, 247–248, 284, 300, 309, 317, 322–325, 328, 331, 334, 340–343, 345, 346, 349–350, 353, 359, 361–362, 365, 370, 374, 391, 395, 398, 411, 426, 441–443, 450–451, 455–456, 460, 462, 464–466, 468–469, 471–472, 474–476, 478, 482–483

- Information Technology Industry Council (ITIC)
 curve 123, 124, 133, 149, 422

- Innovation 161, 405, 422, 473, 474, 476–479, 482

- Inrush current 140, 151

- Insulated gate bipolar transistor (IGBT) 412, 418

- Insulator(s) 9, 10, 13, 177–179, 202, 205
 on transmission lines 178

- Interconnection 29, 159–163, 181, 277, 319, 320, 369, 374, 423, 436, 440, 451, 460, 480, 482

- Intertie(s) 159, 162, 180

- Intuition 1, 17, 41, 47, 89, 99, 102, 108, 115, 116, 148, 189, 197, 258, 296, 300, 339, 363, 399, 405, 473

- Inverter(s) 84, 95, 117, 130, 149, 170–171, 366, 376–377, 390, 412, 414, 415–423, 425, 434, 437–438, 440–441, 444, 447, 449–450, 453–455, 462, 467, 478, 480

- Ionization 5, 10, 177, 192

- Ions 2, 4–6, 8, 11, 22, 143, 406, 408, 444

- Iron losses 205, 212

- Island(ing) 125, 170, 171, 319, 421, 422, 438, 440–442, 449–451, 456

- Iterative solution 327, 345, 354

j

- Jacobian matrix 346, 349, 351, 355–357

- Joule (unit) 3, 14, 403

k

- Kilovolt-ampere (kVA) 82, 84, 85, 118, 183, 201, 214, 219, 221–223, 283, 284, 423
 Kilowatt (kW) 14, 26, 27, 60, 84, 123, 139, 142, 143, 150, 155, 160, 283, 374, 416, 426, 438, 439, 461, 475, 479, 481, 484
 Kilowatt-hour (kWh) 14, 15, 26, 27, 60, 84, 97, 123, 142, 143, 150, 155, 160, 374, 438, 445, 461, 473, 479, 481, 484, 485
 Kirchhoff's current law (KCL) 31, 35–38, 43, 93, 94, 108, 109, 209, 237, 340, 343
 Kirchhoff's laws 29, 35–39, 92–94, 171, 239, 307, 326, 361, 363
 Kirchhoff's voltage law (KVL) 31, 35–38, 92, 138, 216, 217, 366, 367

l

- Lagging current 74, 77, 81, 85, 149, 295, 400, 414, 419
 and power factor 85, 140, 217, 241, 273–275, 281, 285, 295, 400
 Lagrangian 312, 314–317
 Leading current 76, 218, 281, 295, 400
 and power factor 218, 241, 273–276, 285, 289, 295, 399–400, 422
 Light 1, 8, 11, 14–15, 23–25, 61, 70, 122, 135–139, 142, 144, 249, 353, 406, 422, 431, 433–434
 speed of 6–7, 24, 252, 336, 475
 Light emitting diode (LED) 14, 27, 52, 85, 136–138, 143, 154
 Lightning 5, 10, 122, 178, 193, 249, 406, 420
 LinDistFlow 190, 363–368, 422
 Line losses. *See* Losses
 Line drop 61, 186, 188
See also Voltage drop
 Line-to-line and line-to-ground voltage 107–110, 117, 146, 236, 244
See also Phase-to-phase and phase-to-ground
 Linear circuit elements 7, 30, 326
 Linear equations 238, 315, 317, 357, 365
 Lithium-ion batteries 446
 Load
 aggregate 125, 149, 150, 299
 coincident and non-coincident 151
 definition and usage 135–136

Load balancing

- among feeders 174
 among phases 363n38

Load duration curve (LDC) 151–154, 156, 312–314, 458

Load factor 153, 154, 156, 159–161, 201

Load flow 321

See also Power flow

Load-following plant 313

Load forecast(ing) 150, 370, 459, 468

Load frequency control (LFC) 162, 299–311, 319, 456, 466

Load growth 143, 165, 369, 440, 459, 479

Load profile 151–154

Load tap changer (LTC) 187, 205, 376, 439

Looped distribution system 169, 171–173

Loop flow 39, 170–173

Lorentz equation 21

Losses

- and distributed generation 439
 and load balancing 174
 on transmission lines 61–62, 96, 161, 316, 360
 reactive 84, 96, 185, 328–329, 333–335, 350, 353
 resistive 13, 16, 161, 186, 350
 Lossless line 175, 183, 185, 239, 246, 250, 252, 358, 395, 396

Loss-of-load-expectation (LOLE) 371

Loss-of-load-probability (LOLP) 370, 371

Low-pass filter 66, 132, 414

Lumped circuit(s) 7, 251–252

m

Machine, definition and usage 140, 257–258, 262, 286, 300, 384, 436

Magnetic field 21

- of armature 185, 258, 261, 265–268, 271–274, 281, 286–287, 292–293, 382
 of generator 258–261, 264–266, 268–274, 291–293

- of inductor 64–67, 81
 of rotor. *See* Rotor field

Magnetic flux. *See* Flux, magnetic

Magnetic moment 20, 205

Magnetic poles 19, 55, 140, 264, 268, 287, 426
 and rotor windings 268, 272, 291

Magnetization 49, 50, 132, 140, 211–213, 226

- Magnetomotive force (mmf) 50, 203, 292
 Magnetostriction 212, 213
 Majority carrier 410
 Maps 48, 106, 162, 471
 Market(s) 37, 135, 141, 150, 155, 172, 286, 307, 312, 314, 316, 361, 372, 373, 415, 433, 442, 446, 455, 457–460, 468, 480–485
 and power flow 37, 135, 155, 307, 482–484
 Market power 483
 Matrix 34, 38, 114, 115, 238, 239, 241, 243, 250, 317, 323, 340–344, 346, 349–351, 354–359
 Metals 4, 5, 202
 Maximum power point tracking (MPPT) 421
 Mercury arc valve 406–408
 Mho (unit) 9
 Microgrid(s) 126, 159, 161, 171, 296, 302, 373, 415, 421, 441, 444, 449–451, 480
 Microinverter 434
 Microturbines 437–439
 Microwave oven 142
 Minority carrier 410
 Mismatch 85, 302, 328, 345, 349–352, 355, 368, 434
 Modular multi-level converter (MMC) 180
 Monitors, TV and computer 271
 Monopoly 372, 459, 482
 Motor(s)
 efficiency of 141
 energy consumption by 124–125, 141
 induction 85, 139–140, 149, 154, 285. *See also*
 Induction generator
 paper clip 258, 262–264
 single-and three-phase 140–141, 144
 Multijunction PV cells 433
 Mutual inductance 51, 116, 175, 178
- n**
- N-1 criterion 375
 Negative sequence 113–116
 Network reduction 33–34, 44, 45
 Neutral 2–4, 10, 56, 102–104, 106–111, 116–119, 133, 143–146, 187, 201, 206–209, 211, 212, 222, 236, 237, 244, 246, 248, 254, 255, 410, 429, 448, 479
 Conductor 103, 104, 109, 133
 Newton's method 345, 347–349
 Node(s) 38, 94, 270, 321, 322, 324, 325, 340–344, 361–364, 366, 368, 381, 394, 461, 469, 475
 Nonlinear equations 321, 325, 326, 363, 364, 397
 Nonlinearity 325–327, 353
 Nonlinear loads 128, 131, 132, 135, 143
 North American Electric Reliability Corporation (NERC) 162, 467
 Norton equivalent 34, 41–48, 53, 342, 343
 n-type material 410
 Nuclear power 430–432
- o**
- Ocean thermal energy conversion (OTEC) 429
 Oersted, Hans Christian 20
 Ohm (unit) 8, 9, 14, 62, 65, 67, 73, 211, 215, 218, 220, 221, 239, 246, 343
 Ohm's law 7–12, 14–16, 33, 38, 40, 42, 45, 50, 61, 62, 64–66, 77, 83, 86, 91, 92, 111, 116–117, 121, 128, 136, 138, 186, 189, 216, 238, 249, 272, 326, 331, 340, 342, 350, 353, 361, 362, 462
 Oil
 in circuit breakers 194
 as fuel 429, 443
 in transformers 183, 192, 205–206
 One-line diagram 101, 117, 164, 323, 324
 One-port 41–47, 95, 98, 238
 On-load tap changer (OLTC) 187
 Open circuit. *See* Circuit, open
 Open-circuit test 215, 216
 Open-circuit voltage 43, 45, 47, 293, 433, 434
 Operators 124–126, 150, 162, 170, 171, 173, 185, 242, 246, 280, 282, 306, 311, 335, 336, 360, 362, 363, 372, 375–377, 393, 403, 424, 428, 431, 440, 441, 454, 456–458, 461, 462, 466, 468, 469–479, 482, 483
 Optimal power flow (OPF) 316, 322, 360–361
 Oscillations 56, 127, 181, 276, 299, 308, 336, 338, 377, 381, 383, 384, 390, 392, 396, 420, 421, 424, 455, 466, 480
 Outage 110, 151, 161, 171, 173, 174, 313, 350, 363, 370–374, 393, 403, 421, 431, 450, 451, 457, 458, 460, 461, 465, 468, 474, 475, 481
 Outlet 12, 15, 29, 38, 60, 121, 123, 135, 137, 142–146, 192, 311, 416
See also Socket

- Overexcited 295
 Overhead conductors 165, 174, 182
- p**
- Pacific DC Intertie (PDCI) 180
 Parallel connection 29, 30, 33, 35
 Paralleling of generators 280
 Peak factor 200
 Peaking plant 312–313
 Peak load 143, 156, 370, 376, 444, 468
 Penalty factor 314, 316, 320
 Per-unit (p.u.) 86, 175, 211, 215, 218–223, 239, 244, 298, 300, 301, 304, 305, 308, 318, 333, 343, 359, 382, 385, 390, 399, 423
 Period 5, 24, 42, 56, 58, 66, 68, 127, 150, 151, 160, 264, 277, 311, 371, 373, 374, 383, 385, 386, 389, 431, 444
 Permeability 49–50, 63, 204, 211, 226, 228, 230, 264, 269, 292
 constant 49
 Permittivity 235
 Phase
 angle definition 57–59, 321
 conductor 110, 116, 145, 197, 229, 231–232, 234, 236, 238
 constant 249
 shift between voltage and current 74, 78
 spacing 237
 Phase-lock loop (PLL) 420
 Phase-to-phase and phase-to-ground voltage 207
 See also Line-to-line and line-to-ground
 Phasor
 definition of 78, 87–88
 operations with 88, 91
 Phasor diagram 92–94, 98, 99, 104, 107–109, 119, 190, 241, 245, 294–298
 Phasor measurement unit (PMU) 98, 311, 465, 466
 Philosophy, of electric service 370, 372, 374
 Photoelectric effect 25n33
 Photon(s) 25, 26, 42, 43, 142, 433
 Photovoltaics (PV) 170, 432–434, 480
 π -equivalent circuit 242, 245, 251
 Planck's constant 433n15
 Planning 84, 105, 150, 307, 360, 361, 369, 370, 376, 442, 449, 457–460, 480, 481
 Plasma 5, 192, 193, 408
 p-n junction 409–411, 433
 Point of common coupling (PCC) 450
 Point-on-wave (POW) measurements 464
 Polychlorinated biphenyls (PCBs) 206
 Positive sequence 103, 111–113, 115, 117
 Potential 2–4
 See also Voltage
 Power
 active. *See* Power, real
 apparent 79, 81–82, 84–85, 110, 131, 182–183, 189, 219, 244, 283–284, 333–335, 423
 average 62–63, 78–80, 137, 381, 414
 complex 77–86, 94–96, 110–111, 294, 295, 339, 341, 353, 364, 366–367, 394–395, 397
 definition of 13–16, 55, 60–61
 dissipated 56, 61–62
 dissipation in load 138
 instantaneous 42, 62, 77–81, 94, 105, 117, 124, 131, 137, 143, 280, 376, 414, 420, 445
 output from generator 188, 270–271, 277, 279–285, 291, 300–306, 309, 314, 329, 332, 378, 381, 383, 388, 392
 reactive. *See* Reactive power
 real 78–82, 84–85, 96, 117, 125, 131, 140, 182–185, 188–189, 241–242, 270–272, 276–285, 295–296, 321, 327–333, 335, 338, 339, 345, 350, 355, 357–358, 364–365, 378–379, 383, 392–398, 400–402, 422–424, 458, 460, 465, 466
 standby 142
 transmitted on line 16, 61, 185
 true. *See* Power, real
 Power angle 183, 277–280, 330, 377–387, 391–392
 See also Voltage angle
 Power exchange (market) 162, 279, 307, 379, 457, 460
 Power factor
 aggregate 83
 definition of 82, 131
 displacement 78–79, 82–85, 131–132, 143, 217, 241, 398
 distortion 85, 131–132, 143
 of generator 140, 267, 273, 275–276, 282, 284, 285, 289, 291, 295
 and induction generator 285, 289

- Power factor (*contd.*)
 and rotor field 274–276, 281, 285
 significance of 254, 274
 system 83, 132, 140, 276, 282, 285
 true 132n16
 of typical loads 85
- Power flow
 analysis 38, 84, 172, 307, 321–323, 325, 327–331, 336, 338, 339, 342–343, 345, 350, 360–361, 363, 392, 402, 439, 470
 decoupled 355–357
 direction of 332, 360
 equations 339–355, 363
 solution methods 339–355
- Power island. *See Island(ing)*
- Power pool 458
- Power quality 121–134, 137, 141, 143, 155, 171, 371, 415, 421, 443, 451, 464
- Power supplies 135, 142, 321, 372
- Power system(s)
 behavior 81n27, 128, 270, 334, 455
 design and evolution 425, 485. *See also Grid interconnection of* 159–163, 277
 operating state 280, 321, 334, 350, 360, 468
 structure of 159–174
- Power system stabilizer (PSS) 276
- Precision 106, 125, 126, 345, 348, 463, 474–476
- Predictability 420, 439, 477
- Price 141, 155, 316, 443, 481–485
- Primary frequency control 305, 306, 310, 311
- Prime mover 258, 260, 268, 270–272, 283, 285, 288, 296, 297, 300–302, 336, 385, 420, 425
- Propagation constant 248, 251, 252, 255
- Protection
 coordination 194–197
 and distributed generation 440
 zones 194–196
- Protons 2, 4, 19, 20, 409, 410
- p-type material 410
- Public Utility Commission (PUC) 372
- Pulse-width modulation (PWM) 137, 417, 419, 420
- Pumped hydro 426, 443, 444, 454
- q**
- Quadrature 117–119
- r**
- Radians (unit) 56, 58, 252, 351, 384
 per second 58, 384
- Radiation 14, 23–26, 142, 206
- Radio 23–25, 27, 29, 431, 461, 478
- Radioactivity 431
- Ramp rate(s) 135, 153, 154, 299, 306, 313, 431, 457, 458, 482, 483
- Rate of change of frequency (ROCOF) 420, 465
- Rating(s)
 dynamic 182
 emergency 182, 375
 generator 431
 historical trend of 182
 thermal 82, 183, 375, 396
 transmission line 165, 182, 375, 396
- Reactance
 capacitive 63, 67, 69–70, 74, 75, 83–84, 86, 128, 131, 237, 289, 329
 definition of 63, 67
 inductive 56, 63, 67, 69–70, 74–75, 84, 136, 175, 180–182, 211, 217, 231, 234, 240, 253, 273, 279, 281, 289, 358, 392, 396, 398, 422
- Reactive capability curve 284, 285, 297, 423
- Reactive compensation 84–86, 132, 134, 188–190, 334, 424
- Reactive margin 402
- Reactive power
 allocation of 280, 282
 balance of 95
 conservation of 280
 consumed or supplied 281, 328
 definition of 79–80
 direction of flow 333
 and distributed generation 439
 and generators 140, 185, 189, 258, 267, 270, 272–277, 280–285, 289, 295, 327, 329–330, 449
 and motors 140
 and voltage 85, 94–96, 125, 146–149, 181, 185–190, 242, 258, 270, 272–273, 276–277, 280–282, 295, 325, 327–333, 345, 352, 355, 357, 358, 365, 366, 392–394, 396–402, 422–424, 436–437, 439, 449, 484
- Recloser 193–197, 202, 392
- Red loss 433

- Redundancy 168, 169, 194, 362, 363, 370, 371, 375, 477, 478
 Referred impedance 213–215
 Refrigerator 19, 20, 28, 124, 139, 148, 151, 161, 372, 447
 Regulation constant 301, 302, 304, 308, 310, 318
 Relay(s) 110, 125, 142, 166, 170, 191, 192, 196, 202, 306, 310, 393, 440, 451, 455, 471
 Relay desensitization 440
 Reliability 110, 123, 159, 161, 162, 170, 369–374, 416, 432, 450, 451, 467, 473–475
 Reluctance 49–51, 140, 204, 211, 289–290, 476
 Machine 289–290
 Remanence 212, 213
 Remote terminal unit (RTU) 461
 Renewable resources 150, 153, 425, 426, 442, 460, 479, 480, 483, 484
 Reserve(s) 25, 87, 89, 93, 147, 159, 161, 239, 313, 370, 374, 375, 393, 453, 458, 459, 465, 484
 Spinning 458, 484
 Reserve margin 161, 370, 374, 459
 Resistance
 and effect on power consumption 15, 61
 in parallel 31–33, 45–46
 in series 31, 44–46, 138, 463
 Resilience 171, 360, 373, 441, 451, 480
 Resistive heating 13–16, 60, 131, 137, 139, 154, 161, 182, 269
 Resistivity 8, 9, 174, 201, 202, 228
 Restoration 171, 173, 372, 421, 436, 438, 457, 473
 Rheostat 138, 139, 156
 Right-hand rule 20, 22, 63, 209, 230
 Robustness 412, 476, 477
 Root-mean-square (rms) 59, 79, 130, 131, 137, 455, 464, 465
 derivation of 60
 Rotational frequency 260, 268, 270, 271, 277, 384, 387
 Rotor
 cylindrical 265
 round 268, 291–293
 salient pole 268, 291, 292
 squirrel-cage 287
 wound 268, 287
 Rotor current 140, 264, 265, 277, 287, 291, 296, 436
 Rotor field 185, 264, 265, 274–276, 280–282, 293, 295
 in induction machine 285, 287–290
 interaction with stator field 269, 285
 Rotor winding 258, 264, 268, 272, 291, 296
 Rubber bands 3, 263, 338, 339
- S**
- Safety 13, 55, 131, 133, 138, 144, 146, 163, 171, 173, 193, 238, 283, 285, 372, 374, 430, 432, 442, 446, 463, 473–475, 477–479
 Sag 13, 123, 149, 174
 Saturation 129, 212, 213, 405
 Scheduling coordinator 457
 Secondary frequency regulation 303–306, 308, 309, 318
 Security 155, 360, 361, 369, 370, 374–376, 393, 403, 457, 458, 480, 484
 Security-constrained dispatch 458
 Self-excitation 270
 Semiconductors 43, 262, 406, 408, 409
 Series connection 30, 33, 35
 Service 105–106, 121–125, 130, 140–141, 143–145, 149–151, 153–156, 160–163, 165, 167–170, 173, 186, 191, 194, 196, 246, 282, 299, 305–307, 309, 323, 370–375, 385, 423, 438, 441, 447, 451, 458–459, 461, 473, 481, 484
 valuation of 374
 Sheding of load 370, 449
 Shock 1, 10, 12–13, 144, 300, 338, 436, 451
 Short circuit
 current 42–43, 433
 test 215–216
 Siemens (unit) 9, 70, 75, 86, 239, 253, 342, 343
 Silicon 409, 410, 413, 433
 Silicon-controlled rectifier (SCR) 413
 Sine function, definition of 57, 58
 Sinusoidal waveform 78, 79, 86, 126–128, 465
 Siting
 of generators 437, 439
 of transmission lines 369, 459
 Situational awareness 471, 475, 478
 Slack bus. *See* Bus(es), slack
 Slip 139, 262, 286–289, 297, 436
 Socket 144, 146

- Solar power 421, 432, 433, 441
See also Photovoltaics
- Solar storms 133
- Solar thermal power 432, 447
See also Concentrating solar power
- Solenoid 63, 264
- Solid-state technology 179, 262, 272, 405, 408–415, 423, 424
- Source, current or voltage 59, 117, 462
- Source impedance 106, 128–130, 198, 223
- Specific energy and power 446, 473
- Speed
- of light 6–7, 24, 252, 336, 475
 - of signal propagation 7, 250, 252
- Spike 65, 122, 141, 346, 460
- Spin 20, 212, 261, 263, 264, 267, 268, 286
- Spinning reserve. *See* Reserve(s), spinning
- Square wave 416, 419
- Stability
- angle 183–184, 376, 378–392
 - dynamic 338, 377, 382, 384, 385, 435
 - small-signal and large-signal 377
 - steady-state 338, 378–379, 382
 - transient 377, 382, 389
 - voltage 183, 376, 392–394, 399, 455
- Stability limit 180, 183–184, 338, 360, 375, 380, 381, 389, 392, 393, 424
- Starting current 27, 98, 106, 140
- State estimation 361–363
- State estimator 362, 363
- State of charge (SOC) 42, 445, 450
- State variables 358, 362
- Static VAR compensator (SVC) 188, 393, 397
- Stator 139, 140, 264–267, 269, 272–276, 283, 285, 287–292, 436
- Stator field 267, 269, 274–276, 285, 287–289
decomposition of 275
- Steady-state 81, 94, 128, 149, 197, 198, 246, 249, 250, 277, 291, 299, 336, 338, 376–382, 396, 424, 466, 479
- Steam 56, 268, 270, 271, 278, 280, 297, 301, 302, 304, 306, 312, 336, 385, 403, 420, 425, 427–431, 437, 438, 447, 456, 470, 474
- Steam generation 297, 306, 312, 426, 427, 429, 430, 432, 437, 447, 456, 474
- Storage 5, 26, 69, 149, 154, 155, 170, 171, 373, 374, 390, 418, 420, 422, 425, 426, 430, 432, 437, 441–451, 454, 479, 480, 482
- valuation of 374
- String inverter 434
- Substations 164, 166–167, 205, 322, 381, 393, 441, 460
- Subsynchronous oscillations 127, 421
- Subtransmission 164, 167, 168, 322
- Sulfur hexafluoride (SF₆) 192, 206
- Supercapacitors 448
- Supercritical 428
- Superconducting magnetic energy storage (SMES) 5, 448
- Superconducting transmission lines 5, 181–182
- Superconductivity 5, 181
- Superposition principle 39–41
- Supervisory Control and Data Acquisition (SCADA) 173, 461, 462, 478
- Surge impedance 184–185, 202, 246, 248, 252
- Surge impedance loading (SIL) 184–185, 202, 252, 253
- Susceptance 70, 75, 76, 96, 97, 237, 341, 344, 356, 358
- Swell 122
- Swing bus. *See* Bus(es), slack and swing
- Swing equation 382–386, 389–391, 479
- Switch(es) 11, 27, 33, 70, 86, 115, 129, 137–139, 142, 149, 151, 166, 168, 169, 173, 188, 192, 198, 213, 239, 336, 406–409, 411, 412, 415, 416, 418, 419, 424, 428, 440, 449–451, 457, 460, 471, 478
- Switchgear 165, 192, 193, 200
- Switching operations 173, 174, 323, 460, 474, 475
- Symmetrical components 111–117, 119, 197
- Synchronicity of generators 183, 267, 273, 277, 278, 280, 299, 327, 328, 396, 424
- Synchronism 183, 378, 379
- Synchronization 126, 280, 287, 288, 297, 298, 304, 318, 378, 379, 381, 384, 390, 425, 438, 439, 453, 479
- Synchronous generator 125, 140, 257, 258, 264–270, 272, 277, 280, 287, 288, 297, 298, 304, 318, 378, 379, 384, 390, 425, 438, 439, 453, 479
- Synchronous reactance 294, 297, 298
- Synchronous speed 280, 286, 288–290, 300, 379, 385, 390, 403, 404, 426, 458
- Synchrophasors 465
- Synchroscope 280n20

- System operator 121, 125, 126, 136, 150, 152, 153, 162, 185, 306, 335, 336, 360, 370, 372, 424, 428, 440, 441, 456–458, 462, 466, 469, 482, 483
See also Independent system operator
- t**
- Taylor series 346–348
- Telephone 173, 306, 318, 457, 460, 461, 474, 478
- Tension 2, 178, 338
See also Voltage
- Terminal(s) 4, 6, 11, 15, 29–32, 39, 41, 42, 44–47, 53, 62, 144–146, 180, 185, 222, 262, 263, 269, 270, 272, 286, 289, 291, 293–297, 383, 406–408, 412, 413, 434, 444, 445, 461, 478
- Territory(ies) 72, 150, 162, 166, 167, 369, 371, 374, 458, 459, 468, 471
- Tertiary frequency control 299, 306
- Tesla (unit) 21, 26, 176, 177, 289, 290
- Tesla coil 176, 177
- Tesla, Nikola 289, 290
- Thermal limit 182–184, 205, 277, 335, 338, 394, 400, 424
See also Ratings
- Thermodynamics
- first law 24
 - second law 428
- Thermostatically controlled loads (TCL) 124, 155, 447
- Thévenin equivalent 43–45
- Three-phase
- generator windings 265, 267
 - motors 140–141, 144
 - transmission 101, 110, 117, 144, 232, 234, 236, 244, 254
- Thyristors 180, 412–415, 418, 419
- Tidal power 426
- Tie line flow 307–311, 319
- Time
- as angle 57–58
 - conceptualization of 336
 - and system frequency 454
 - as a variable in a.c. circuits 325
- Time-current curve 191, 195, 196
- Time scale(s)
- for analysis 86
 - for grid operation 454–455
- Time-of-use (TOU) rates 155, 441, 482
- Topping cycle 429
- Torque
- for induction machine 285–287
 - on generator rotor 105, 265
- Total harmonic distortion (THD) 130, 131, 134, 416
- Transformer(s)
- bank 110, 166, 206
 - capacity limit 205
 - cooling 205–206
 - core 129, 183, 203–205, 209, 212, 417
 - distribution 34, 145, 165, 205
 - harmonics 132–133
 - and need for a.c. 55, 203, 405, 415
 - primary and secondary side 145, 204, 206–209, 211, 214, 219, 221–222
 - saturation. *See* Saturation
 - tap(s) 145, 187, 205, 393, 460
 - turns ratio 188, 222
- Transient stability. *See* Stability, transient
- Transistor(s) 30, 128, 142, 408, 409, 411–413, 418, 419, 423
- Transmission line(s)
- capacitance 68, 84, 86, 175, 188, 225, 234–238, 245–246, 344
 - capacity 174
 - clearance 178
 - dimensions of 175
 - inductance 175
 - parameters 457
 - reactance 86
 - sagging. *See* Sag
- Transmission system 56, 106, 161, 163, 164, 167, 168, 193, 280, 321, 322, 333, 369, 370, 455, 457, 468
- Transmission system operator (TSO) 457
- Transparency 475–477
- Transposition 178, 179
- Triplen harmonics 132–134
- Turbine 258, 260, 268, 270, 271, 277, 278, 280, 283, 286, 297, 300, 318, 379, 382, 383, 385, 386, 389, 403, 404, 415, 425–432, 435, 436, 443, 444, 461, 470
- Two-port 238–239, 250

u

- Underexcited 289, 295
 Underground 123, 165, 174, 175, 182, 371, 373, 430, 447
 Uninterruptible power supply (UPS) 372, 448, 450, 451, 476
 Unit commitment 457, 458
 Unsymmetrical fault 197–202
 Utility(ies) 12, 15, 60, 83, 84, 86, 98, 106, 109, 121–123, 128, 132, 137, 140, 141, 143–146, 150, 151, 153–155, 160–162, 164, 170, 171, 173, 177, 183, 189, 190, 197, 206, 264, 319, 369, 371–374, 421, 423, 433, 438, 440, 441, 444, 447, 450, 451, 457–462, 468–470, 474, 478, 481, 482, 485

v

- VAR compensation 188, 393, 397
See also Reactive compensation; Reactive power
 Variables in power flow 331
 Variable speed drive (VSD) 85, 141
 Vector(s) 21–22, 38, 50, 73–74, 81–82, 107–108, 112–115, 117, 216, 225–226, 230, 265–267, 289, 349, 351, 362
 addition 92, 113, 217, 265
 Veracity 476, 477
 Volt (unit) 3, 8, 80, 180, 182, 253, 366, 421–423, 439
 Voltage
 control 122, 124, 185–190, 329, 393, 400, 423
 dependence of loads 394
 at generator terminal 185, 272, 288, 291, 294
 tolerance 123
 Voltage angle 200–201, 325, 328, 330–333, 335–336, 338, 345, 352, 355–360, 365, 390, 393, 397, 399
 in generator 277, 330, 378, 383. *See also* Power angle
 Voltage collapse 328, 393, 403
 Voltage divider 209, 463, 464
 Voltage drop 11–14, 16, 31–33, 35, 38, 43, 46, 47, 64, 65, 77, 85, 92, 93, 99, 104, 106, 116, 121, 122, 129, 130, 138, 140, 149, 186, 187, 189, 190, 200, 201, 215–218, 220, 240–245, 247, 254, 276, 294, 295, 326, 353, 363–366, 395, 399, 400, 402, 449

- Voltage level 10, 12, 16, 39, 62, 85, 109, 110, 122, 137, 141, 145, 164, 179, 182, 185, 186, 205, 218, 221, 269, 281, 360, 393, 396, 422, 445, 449, 456, 478

Voltage profile 123, 187, 189, 281, 354, 356, 357, 422, 439, 458

Voltage regulation 85, 143, 149, 216–218, 220, 223, 244, 329, 422

Voltage regulator 168, 187, 188, 276, 282, 285, 329, 393, 439, 449, 460, 478

Voltage ride-through 421

Voltage sag 122, 123, 149, 198

Voltage source converter (VSC) 412, 413, 415, 418

Voltage spike 65

Voltage stability 183, 376, 392–394, 399, 455

Voltage swell 122

Volt-ampere (VA) 80–82, 182, 253, 283, 385

See also Apparent power

Volt-ampere reactive (VAR) 80

See also Reactive power

Volt-VAR control 422

Volt-watt control 421

Vulnerability 105, 123, 161, 210, 283, 372, 374, 415, 441, 479

w

Water flow analogy 37, 271

Watt (unit) 14, 15, 60, 271, 276, 301, 366, 421–423, 446, 447

Watt-hour (unit) 14, 15, 26, 27, 60, 84, 123, 142, 150, 155, 160, 314, 374, 443, 444, 446, 447, 461, 473, 479, 481, 483, 484

Waveform 78, 82, 86, 98, 121, 126–134, 137, 181, 201, 212, 265, 381, 405, 413–416, 418–420, 424, 462, 464–467, 479

Wave, propagation of 24, 246

Weber (unit) 21

Western Electric Coordinating Council (WECC) 162

Westinghouse, George 55, 405

Wild a.c. 415, 437

Wind power 286, 312, 420, 435–437, 442–443, 480, 485

and generator rating 436

- Work 3, 5, 11–14, 17, 42, 48, 51, 62, 66, 71, 78, 85, 101, 103, 116–117, 124, 131, 135, 138, 142–143, 153, 173, 174, 197, 200, 205, 212–213, 218, 237, 248, 258, 260, 262, 285, 290, 300, 305, 324–325, 331, 338, 340, 345, 350, 387–389, 412, 414, 428, 433, 438, 444–445, 448, 453–485
- Wye connection 106–109
grounding of 104, 110

Z

- Zero-sequence 111, 113, 115, 116, 119, 133, 197
- ZIP loads 124, 147