

Contents

Part I Standard Representations

1	Number Systems	3
1.1	Representing Numbers	3
1.2	The Big Three (and One Old Guy)	8
1.3	Converting Between Number Bases	10
1.4	Chapter Summary	16
	Exercises	16
	References	17
2	Integers	19
2.1	Bits, Nibbles, Bytes, and Words	19
2.2	Unsigned Integers	21
2.2.1	Representation	21
2.2.2	Storage in Memory: Endianness	22
2.3	Operations on Unsigned Integers	25
2.3.1	Bitwise Logical Operations	25
2.3.2	Testing, Setting, Clearing, and Toggling Bits	30
2.3.3	Shifts and Rotates	33
2.3.4	Comparisons	37
2.3.5	Arithmetic	41
2.3.6	Square Roots	52
2.4	What About Negative Integers?	54
2.4.1	Sign-Magnitude	54
2.4.2	One's Complement	55
2.4.3	Two's Complement	55
2.5	Operations on Signed Integers	56
2.5.1	Comparison	56
2.5.2	Arithmetic	58
2.6	Binary-Coded Decimal	67
2.6.1	Introduction	67
2.6.2	Arithmetic with BCD	69

2.6.3	Conversion Routines	70
2.6.4	Other BCD Encodings	73
2.7	Chapter Summary	76
	Exercises	77
	References	79
3	Floating Point	81
3.1	Floating-Point Numbers	81
3.2	An Exceedingly Brief History of Floating-Point Numbers	84
3.3	Comparing Floating-Point Representations	85
3.4	IEEE 754 Floating-Point Representations	89
3.5	Rounding Floating-Point Numbers (IEEE 754)	97
3.6	Comparing Floating-Point Numbers (IEEE 754)	100
3.7	Basic Arithmetic (IEEE 754)	102
3.8	Handling Exceptions (IEEE 754)	105
3.9	Floating-Point Hardware (IEEE 754)	108
3.10	Binary Coded Decimal Floating-Point Numbers	110
3.11	Chapter Summary	113
	Exercises	114
	References	115
4	Pitfalls of Floating-Point Numbers (and How to Avoid Them)	117
4.1	What Pitfalls?	117
4.2	Some Experiments	119
4.3	Avoiding the Pitfalls	130
4.4	Chapter Summary	134
	Exercises	135
	References	135
 Part II Other Representations		
5	Big Integers and Rational Arithmetic	139
5.1	What is a Big Integer?	139
5.2	Representing Big Integers	140
5.3	Arithmetic with Big Integers	146
5.4	Alternative Multiplication and Division Routines	158
5.5	Implementations	167
5.6	Rational Arithmetic with Big Integers	171
5.7	When to Use Big Integers and Rational Arithmetic	177
5.8	Chapter Summary	180
	Exercises	180
	References	181
6	Fixed-Point Numbers	183
6.1	Representation (Q Notation)	183
6.2	Arithmetic with Fixed-Point Numbers	188
6.3	Trigonometric and Other Functions	194

6.4	An Emerging Use Case	204
6.5	When to Use Fixed-Point Numbers	211
6.6	Chapter Summary	212
	Exercises	212
	References	213
7	Decimal Floating Point	215
7.1	What is Decimal Floating-Point?	215
7.2	The IEEE 754-2008 Decimal Floating-Point Format	216
7.3	Decimal Floating-Point in Software	225
7.4	Thoughts on Decimal Floating-Point	232
7.5	Chapter Summary	233
	Exercises	234
	References	234
8	Interval Arithmetic	235
8.1	Defining Intervals	235
8.2	Basic Operations	237
8.3	Functions and Intervals	253
8.4	Implementations	258
8.5	Thoughts on Interval Arithmetic	262
8.6	Chapter Summary	263
	Exercises	263
	References	263
9	Arbitrary Precision Floating-Point	265
9.1	What is Arbitrary Precision Floating-Point?	265
9.2	Representing Arbitrary Precision Floating-Point Numbers	265
9.3	Basic Arithmetic with Arbitrary Precision Floating-Point Numbers	270
9.4	Comparison and Other Methods	273
9.5	Trigonometric and Transcendental Functions	274
9.6	Arbitrary Precision Floating-Point Libraries	278
9.7	Thoughts on Arbitrary Precision Floating-Point	290
9.8	Chapter Summary	291
	Exercises	291
	References	292
10	Other Number Systems	293
10.1	Introduction	293
10.2	Logarithmic Number System	293
10.3	Double-Base Number System	307
10.4	Residue Number System	324
10.5	Redundant Signed-Digit Number System	332
10.6	Chapter Summary	339
	Exercises	340
	References	341
	Index	343

Numbers and Computers

Kneusel, R.T.

2017, XIII, 346 p. 68 illus., 12 illus. in color., Hardcover

ISBN: 978-3-319-50507-7