Computer Architecture–Homework VII

107 Fall semester, Chapter 10

10.25 Consider a reduced 7-bit IEEE floating-point format, with 3 bits for the exponent and 3 bits for the significand. List all 127 values.

Sol:

sign_bit=0⇒positive numbers	zero & subnormal	normal nonzero						infinity & NaN
		2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	
significand ↓ \ exponent →	000	001	010	011	100	101	110	111
$(1)_{10} = (1.000)_2 \Leftarrow 000$	0	0.25	0.5	1	2	4	8	∞
$(1.125)_{10} = (1.001)_2 \Leftarrow 001$	0.03125	0.28125	0.5625	1.125	2.25	4.5	9	
$(1.25)_{10} = (1.010)_2 \Leftarrow 010$	0.0625	0.3125	0.625	1.25	2.5	5	10	
$(1.375)_{10} = (1.011)_2 \Leftarrow 011$	0.09375	0.34375	0.6875	1.375	2.75	5.5	11	
$(1.5)_{10} = (1.100)_2 \Leftarrow 100$	0.125	0.375	0.75	1.5	3	6	12	NaN
$(1.625)_{10} = (1.101)_2 \Leftarrow 101$	0.15625	0.40625	0.8125	1.625	3.25	6.5	13	
$(1.75)_{10} = (1.110)_2 \Leftarrow 110$	0.1875	0.4375	0.875	1.75	3.5	7	14	
$(1.875)_{10} = (1.111)_2 \Leftarrow 111$	0.21875	0.46875	0.9375	1.875	3.75	7.5	15	

sign_bit=1⇒negative numbers	zero & subnormal	normal nonzero						infinity & NaN
		2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	
significand ↓ \ exponent →	000	001	010	011	100	101	110	111
$(1)_{10} = (1.000)_2 \Leftarrow 000$	-0	-0.25	-0.5	-1	-2	-4	-8	$-\infty$
$(1.125)_{10} = (1.001)_2 \Leftarrow 001$	-0.03125	-0.28125	-0.5625	-1.125	-2.25	-4.5	-9	
$(1.25)_{10} = (1.010)_2 \Leftarrow 010$	-0.0625	-0.3125	-0.625	-1.25	-2.5	-5	-10	
$(1.375)_{10} = (1.011)_2 \Leftarrow 011$	-0.09375	-0.34375	-0.6875	-1.375	-2.75	-5.5	-11	
$(1.5)_{10} = (1.100)_2 \Leftarrow 100$	-0.125	-0.375	-0.75	-1.5	-3	-6	-12	NaN
$(1.625)_{10} = (1.101)_2 \Leftarrow 101$	-0.15625	-0.40625	-0.8125	-1.625	-3.25	-6.5	-13	
$(1.75)_{10} = (1.110)_2 \Leftarrow 110$	-0.1875	-0.4375	-0.875	-1.75	-3.5	-7	-14	
$(1.875)_{10} = (1.111)_2 \Leftarrow 111$	-0.21875	-0.46875	-0.9375	-1.875	-3.75	-7.5	-15	

For better understanding, values can also be represented as fractions.

sign_bit=0⇒positive numbers	zero & subnormal	normal nonzero						infinity & NaN
		2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	
significand ↓ \ exponent →	000	001	010	011	100	101	110	111
$(8/8)_{10} = (1.000)_2 \Leftarrow 000$	0	8/32	8/16	8/8	8/4	8/2	8	∞
$(9/8)_{10} = (1.001)_2 \Leftarrow 001$	1/32	9/32	9/16	9/8	9/4	9/2	9	
$(10/8)_{10} = (1.010)_2 \Leftarrow 010$	2/32	10/32	10/16	10/8	10/4	10/2	10	
$(11/8)_{10} = (1.011)_2 \Leftarrow 011$	3/32	11/32	11/16	11/8	11/4	11/2	11	
$(12/8)_{10} = (1.100)_2 \Leftarrow 100$	4/32	12/32	12/16	12/8	12/4	12/2	12	NaN
$(13/8)_{10} = (1.101)_2 \Leftarrow 101$	5/32	13/32	13/16	13/8	13/4	13/2	13	
$(14/8)_{10} = (1.110)_2 \Leftarrow 110$	6/32	14/32	14/16	14/8	14/4	14/2	14	
$(15/8)_{10} = (1.111)_2 \Leftarrow 111$	7/32	15/32	15/16	15/8	15/4	15/2	15	

sign_bit=1⇒negative numbers	zero & subnormal	normal nonzero						infinity & NaN
		2^{-2}	2^{-1}	2^{0}	2^{1}	2^{2}	2^{3}	
significand ↓ \ exponent →	000	001	010	011	100	101	110	111
$(8/8)_{10} = (1.000)_2 \Leftarrow 000$	-0	-8/32	-8/16	-8/8	-8/4	-8/2	-8	$-\infty$
$(9/8)_{10} = (1.001)_2 \Leftarrow 001$	-1/32	-9/32	-9/16	-9/8	-9/4	-9/2	-9	
$(10/8)_{10} = (1.010)_2 \Leftarrow 010$	-2/32	-10/32	-10/16	-10/8	-10/4	-10/2	-10	
$(11/8)_{10} = (1.011)_2 \Leftarrow 011$	-3/32	-11/32	-11/16	-11/8	-11/4	-11/2	-11	
$(12/8)_{10} = (1.100)_2 \Leftarrow 100$	-4/32	-12/32	-12/16	-12/8	-12/4	-12/2	-12	NaN
$(13/8)_{10} = (1.101)_2 \Leftarrow 101$	-5/32	-13/32	-13/16	-13/8	-13/4	-13/2	-13	
$(14/8)_{10} = (1.110)_2 \Leftarrow 110$	-6/32	-14/32	-14/16	-14/8	-14/4	-14/2	-14	
$(15/8)_{10} = (1.111)_2 \Leftarrow 111$	-7/32	-15/32	-15/16	-15/8	-15/4	-15/2	-15	

10.22 Assume that the exponent e is constrained to lie in the range $0 \ldots e \ldots X$, with a bias of q, that the base is b, and that the significand is p digits in length.

(a) What are the largest and smallest positive values that can be written?

(b) What are the largest and smallest positive values that can be written as normalized floating-point numbers?

Sol:

Now a number is represented as $0.\underbrace{b_1 b_2 b_3 b_4 b_5 b_6 \cdots b_p}_{p \text{ digits significand}} \times b^{e-q}$, where b_i is the i-th digit of the significand.

(a) The largest value is $(1 - b^{-p}) \times b^{x-q}$ $\quad \forall i \in \{1, \ldots, p\}, b_i = b - 1$
The smallest value is $b^{-p} \times b^{-q}$ $\quad \forall i \in \{1, \ldots, p-1\}, b_i = 0; b_p = 1$

(b) The largest value is $(1 - b^{-p}) \times b^{x-q}$ $\quad \forall i \in \{1, \ldots, p\}, b_i = b - 1$
The smallest value is $b^{-1} \times b^{-q}$ $\quad \forall i \in \{2, \ldots, p\}, b_i = 0; b_1 = 1$