2.5.4 The IEEE-754 Floating-Point Representation

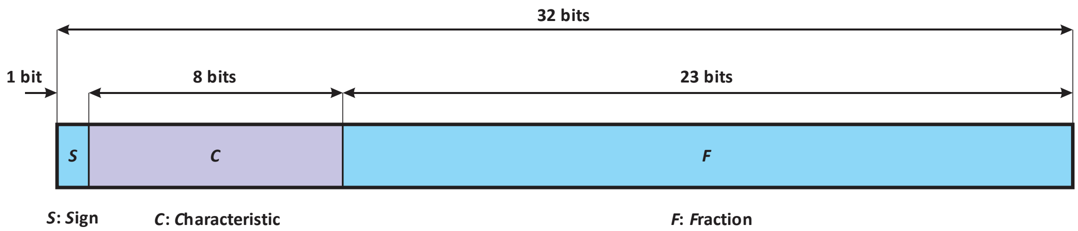

Figure 1. IEEE 754 single binary floating-point format.

Symbol	Name	Description
S	Sign	$0 \Rightarrow$positive; $1 \Rightarrow$negative
B	Bias	The bias is used to convert the characteristic to an unsigned exponent. $C = E + B$ where E is a binary exponent and $B = 2^{k-1} - 1$ and k is the number of bits in the characteristic. Computing B for the IEEE 754 single binary floating-point format, we find $B = 127$.
C	Characteristic	An unsigned binary exponent representing the value 2^{C-B}.
F	Fraction	A binary value equaling $1.F$.

Procedure for formatting an IEEE 754 single binary floating-point number.

Given: V: a rational value
Wanted: 1. S, sign, 1 bit
2. C, characteristic, 8 bits
3. F, fraction, 23 bits
Note: B, the bias, for a single floating-point number is $B = 2^{8-1} - 1 = 127$
Define: $V = (-1)^S \times 2^{C-B} \times 1.F$

1. Find the sign, S.
   1.1. Extract the sign of V. Convert the sign to binary. $+ \Rightarrow 0, - \Rightarrow 1$
2. Find the characteristic, C.
   2.1. Find a power of 2, E, such that
      2.1.1.$1 \le \frac{V}{2^E} < 2$
      2.1.2.$2^E \le V < 2^{E+1}$
      2.1.3.$\log_2 2^E \le \log_2 V < \log_2 2^{E+1}$
      2.1.4.$E = \lfloor \log_2 V \rfloor$
   2.2. Compute C, the characteristic, $C = E + B$
   2.3. Convert $C$to an unsigned 8-bit binary number
3. Find the fraction, F.
   3.1. Find $\frac{V}{2^E} - 1$
   3.2. Convert the fraction to a 23-bit binary value
4. Put components in the IEEE 754 single binary floating point format.
   4.1. Put the sign, S, in bit-position 0.

4.2. Put the characteristic, C, in bit-positions 1 – 8.
4.3. Put the fraction, F, in bit positions 9 – 31.

Example: $V = 67.875$

1. Find the sign, S.
   1.1. Extract the sign of V. Convert the sign to binary. $+ \Rightarrow 0, - \Rightarrow 1$
   1.2. $sign(V) = + \Rightarrow \mathbf{0}$
2. Find the characteristic, C.
   2.1. Find a power of 2, E, such that
      2.1.1. $1 \leq \frac{V}{2^E} < 2$
      2.1.2. $2^E \leq V < 2^{E+1}$
      2.1.3. $\log_2 2^E \leq \log_2 V < \log_2 2^{E+1}$
      2.1.4. $E = \lfloor \log_2 V \rfloor$
      2.1.5. $E = \lfloor \log_2 67.875 \rfloor = \left\lfloor \frac{\log_{10} 67.875}{\log_{10} 2} \right\rfloor = \lfloor 6.084808388 \ldots \rfloor = \mathbf{6}$
   2.2. Compute C, the characteristic, $C = E + B$
   2.3. $C = 6 + 127 = \mathbf{133}$
   2.4. Convert C to an unsigned 8-bit binary number

Divisor	**Dividend**	**Quotient**	**Remainder**	**Foreign Digit**
16	133	8	5	5
16	8	0	8	8

   2.5. $C = 133_{10} = 85_{16} = \mathbf{1000\ 0101_2}$
3. Find the fraction, F.
   3.1. Find $\frac{V}{2^E} - 1$
   3.2. $\frac{V}{2^E} - 1 = \frac{67.875}{2^6} - 1 = \frac{67.875}{64} - 1 = 1.060546875 - 1 = 0.060546875$
   3.3. $F = \mathbf{0.060546875}$
   3.4. Convert the fraction to a 23-bit binary value

Multiplier	**Multiplicand**	**Product**	**Integer Portion**	**Foreign Digit**
16	0.060546875	0.968750000	0.000000000	0
16	0.968750000	15.500000000	15.000000000	F
16	0.500000000	8.000000000	8.000000000	8
16	0.000000000	0.000000000	0.000000000	0
16	0.000000000	0.000000000	0.000000000	0

   3.5. $F = 0.0F800 = 0.\mathbf{0000\ 1111\ 1000\ 0000\ 0000\ 000}$

4. Put components in the IEEE 754 single binary floating-point format.
    4.1. Put the sign, S, in bit-position 0.

S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0																															

    4.2. Put the characteristic, C, in bit-positions 1 – 8.

S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0	1	0	0	0	0	1	0	1																							

    4.3. Put the fraction, F, in bit positions 9 – 31.

S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5. Now, find the hexadecimal value

S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0	1	0	0	0	0	1	0	1	0	0	0	0	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4				2				8				7				C				0				0				0			

$V = 4287C000$

Procedure for converting an IEEE 754 single binary floating-point number in hexadecimal format to scientific notation.

Given: H: a 32-bit hexadecimal number
Wanted $V = (-1)^S \times 2^{C-B} \times 1.F = r \times 10^E$
Note: B, the bias, for a single floating-point number is $B = 2^{8-1} - 1 = 127$

1. Extract the sign, S, from bit 0 and find the sign.
2. Extract the characteristic, C, from bits 1 – 8, convert to decimal, and compute 2^{C-B}
    2.1. Extract the characteristic, C, from bits 1 – 8.
    2.2. Convert to decimal.
3. Extract the fraction, F, from bits 9 – 31, covert to a decimal fraction.
    3.1. Extract the fraction, F, from bits 9 – 31.
    3.2. Covert to a decimal fraction.
4. Find the product of steps 1 – 3.
    4.1. $V = (-1)^S \times 2^{C-B} \times 1.F$

Example:
H=203D26D1

1. Extract the sign, S, from bit 0 and find the sign, $S = \mathbf{0}$

2				0				3				D				2				6				D				1			
0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	0	0	0	1
S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0																															

2. Extract the characteristic, C, from bits 1 – 8, convert to decimal, and compute 2^{C-B}

2				0				3				D				2				6				D				1			
0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	0	0	0	1
S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0	0	1	0	0	0	0	0	0																							

   2.1. Extract the characteristic, C, from bits 1 – 8, $C = 0100\ 0000$

   2.2. Convert to decimal, $C = 0100\ 0000_2 = 40_{16} = \mathbf{64_{10}}$

3. Extract the fraction, F, from bits 9 – 31, covert to a decimal fraction.

2				0				3				D				2				6				D				1			
0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	0	0	0	1
S	C								F																						
0	1	2	3	4	5	6	7	8	9	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2	2	3	3
										0	1	2	3	4	5	6	7	8	9	0	1	2	3	4	5	6	7	8	9	0	1
0	0	1	0	0	0	0	0	0	0	1	1	1	1	0	1	0	0	1	0	0	1	1	0	1	1	0	1	0	0	0	1

   3.1. Extract the fraction, F, from bits 9 – 31, $F = 0111\ 1010\ 0100\ 1101\ 1010\ 001$

   3.2. Covert to a decimal fraction, $F = 0.0111\ 1010\ 0100\ 1101\ 1010\ 001\ _2 = 0.7A4DA2_{16}$

Digit	Decimal Equivalent	Position	Base	Exponentiated	Subtotal
7	7	-1	16	0.0625	0.4375
A	10	-2	16	0.00390625	0.0390625
4	4	-3	16	0.000244141	0.000976563
D	13	-4	16	1.52588E-05	0.000198364
A	10	-5	16	9.53674E-07	9.53674E-06
2	2	-6	16	5.96046E-08	1.19209E-07
Total					**0.477747083**

4. Find the product of steps 1 – 3.

   4.1. $V = (-1)^S \times 2^{C-B} \times 1.F = (-1)^{\mathbf{0}} \times 2^{\mathbf{64}-127} \times 1.\mathbf{477747083}$
   $= \mathbf{1} \times \mathbf{2^{-63}} \times \mathbf{1.477747083} = \mathbf{1.602176597 \times 10^{-19}}$
   the charge on an electron.

Practice problems:

1. Convert Amadeo Avogadro's number, $6.022 \times 10^{23}$to IEEE 754 single binary floating-point representation and express the result as a sequence of hexadecimal digits.
2. Convert **4d8ef3c2,** an IEEE 754 single binary floating-point number to scientific notation.