Example 9.5 **Problem:** Consider the following 12-bit floating-point number representation format that is manageable for working through numerical exercises. The first bit is the sign of the number. The next five bits represent an excess-15 exponent for the scale factor, which has an implied base of 2. The last six bits represent the fractional part of the mantissa, which has an implied 1 to the left of the binary point.

Perform Subtract and Multiply operations on the operands

$$A = \begin{array}{|c|c|c|}\hline 0 & 10001 & 011011 \\ \hline \end{array}$$

$$B = \begin{array}{|c|c|c|}\hline 1 & 01111 & 101010 \\ \hline \end{array}$$

which represent the numbers

$$A = 1.011011 \times 2^2$$

and

$$B = -1.101010 \times 2^0$$

Solution: The required operations are performed as follows:

- Subtraction

  According to the Add/Subtract rule in Section 9.7.1, we perform the following four steps:

  1. Shift the mantissa of B to the right by two bit positions, giving 0.01101010.
  2. Set the exponent of the result to 10001.
  3. Subtract the mantissa of B from the mantissa of A by adding mantissas, because B is negative, giving

     $$\begin{array}{rcccccccccc} & 1 & . & 0 & 1 & 1 & 0 & 1 & 1 & 0 & 0 \\ + & 0 & . & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 0 \\ \hline & 1 & . & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 0 \end{array}$$

     and set the sign of the result to 0 (positive).
  4. The result is in normalized form, but the fractional part of the mantissa needs to be truncated to six bits. If this is done by rounding, the two bits to be removed represent the tie case, so we round to the nearest even number by adding 1, obtaining a result mantissa of 1.110110. The answer is

     $$A - B = \begin{array}{|c|c|c|}\hline 0 & 10001 & 110110 \\ \hline \end{array}$$

- Multiplication

  According to the Multiplication rule in Section 9.7.1, we perform the following three steps:

  1. Add the exponents and subtract 15 to obtain 10001 as the exponent of the result.
  2. Multiply mantissas to obtain 10.010110101110 as the mantissa of the result. The sign of the result is set to 1 (negative).
  3. Normalize the resulting mantissa by shifting it to the right by one bit position. Then add 1 to the exponent to obtain 10010 as the exponent of the result. Truncate the mantissa fraction to six bits by rounding to obtain the answer

     $$A \times B = \begin{array}{|c|c|c|}\hline 0 & 10010 & 001011 \\ \hline \end{array}$$

EXAMPLE 9.1

Form the single-precision representation of 6.5.

Solution

The sign is positive, so the sign bit will be 0. The power of 2 that will result in a significand between 1 and almost 2 is 4.0 (2^2), resulting in a significand of 1.625. Expressed in floating-point representation, the value 6.5 is

$$6.5 = -1^0 \times 2^2 \times 1.625$$

EXAMPLE 9.2

Form the single-precision representation of –0.4375.

Solution

The sign is negative, so the sign bit will be 1. The power of 2 that will result in a significand between 1 and almost 2 is 2^{-2} (0.25), giving a significand of 1.75.

$$-0.4375 = -1^1 \times 2^{-2} \times 1.75$$

$$1.75 = 1 + \tfrac{1}{2} + \tfrac{1}{4}\text{, or in binary, } 1.11.$$

The exponent is –2, and when the bias is added to form the exponent of the single-precision representation, the biased exponent becomes 125, or 0x7D. The resulting single-precision value is 0xBEE00000. See Figure 9.7.

S	Exponent								Fraction																						
3	3	2	2	2	2	2	2	2	2	2	2	1	1	1	1	1	1	1	1	1	1	9	8	7	6	5	4	3	2	1	0
1	0	9	8	7	6	5	4	3	2	1	0	9	8	7	6	5	4	3	2	1	0										
1	0	1	1	1	1	1	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
B				E				E				0				0				0				0				0			

FIGURE 9.7 Result of Example 9.2.

EXAMPLE 11.2

Transfer the flag bits in the FPSCR to the APSR.

Solution

The transfer is made with a VMRS instruction, with the destination APSR_nzcv:

```
VMRS.F32 APSR_nzcv, FPSCR
```