Floating Point Addition (Binary)

1.Align binary points:

- Align binary point of the number with smaller exponent

2. Add significands:

3. Normalize result:

4. Round and renormalize if necessary:

NB. We assumed that significands can be only 4 bits or digits.

repeated Multiplication by 2

0.5

$0.5 \times 2 = 1.0_1$

$0.5_{10} = 0.1_2$

Problem: Perform binary floating-point addition and convert the resulting values to IEEE-754 single and IEEE-754 double precision format. Finally convert them to hexadecimal values. 0.5 + 0.4375

$0.25 + 0.125 + 0.0625 = 0.4375$

Binary point

0.5_{10}

Weight :	0.5	0.25	0.125
	1	0	0

$0.5_{10} = 0.1_2$

$= 1.0 \times 2^{-1}$

Normalized

Weight	0.5	0.25	0.125	0.0625
	0	1	1	1

$0.4375_{10} = 0.0111_2$

$= 1.11 \times 2^{-2}$

Normalized.

1. Align binary points

$0.5 = 1.000 \times 2^{-1}$

$0.4375 = 0.111 \times 2^{-1}$

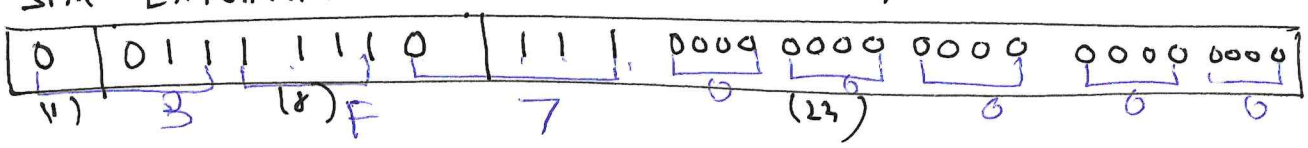
significands.

2. Add significands

$$\begin{array}{r} 1.000 \\ 0.111 \\ \hline 1.111 \end{array}$$

3. Result $= 1.111 \times 2^{-1}$

IEEE 754 single precision Format



$S = 0$

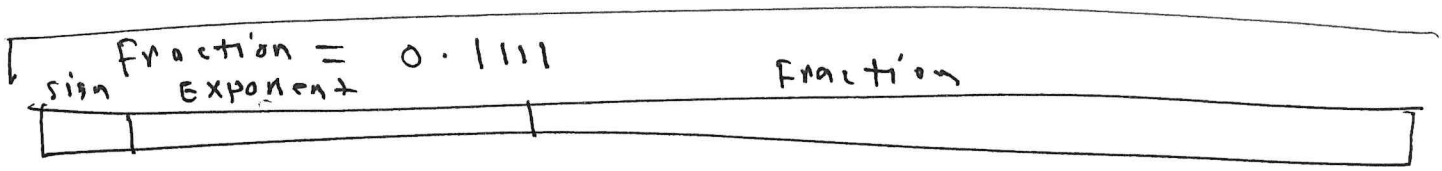
exponent $= -1 + 127 = 126_{10} = 0111\ 1110_2$

Fraction $= 0.11100\ldots$

Hex value in single precision $= 0X3F700000$

IEEE - 754 Double precision Format

$S = 0$, exponent $= -1 + 1023 = 1022_{10}$ $0\ 1111\ 1111\ 10_2$



Floating Point Subtraction (Binary)

1.Align binary points:

- Align binary point of the number with smaller exponent

2. Add significands:

3. Normalize result:

4. Round and renormalize if necessary:

NB. We assumed that significands can be only 4 bits or digits.

Problem: Perform binary floating-point subtraction and convert the resulting values to IEEE-754 single. Finally convert them to hexadecimal values.

0.5 – 0.4375

0.5

1.0×2^{-1}

-0.4375

-1.11×2^{-2}

8

+ (-3)

5.

8

− 3

5

8 8

2's complement of (+3) = -3

5.

+8 = 1000

+3 = 0011

1's complement = 1100

+1

-3 = 2's complement = 1101
of (+3)

+8 = 1000

-3 = 1101

1 0101

5

1. Align Binary points

$0.5 = 1.000 \times 2^{-1}$

$-0.4375 = -0.111 \times 2^{-1}$

2. Add significands

2's complement of $0.111 = 1.000 + 1 = 1.001$

$$\begin{array}{r} 1.000 \\ 1.001 \\ \hline 10.001 \end{array} \times 2^{-1}$$

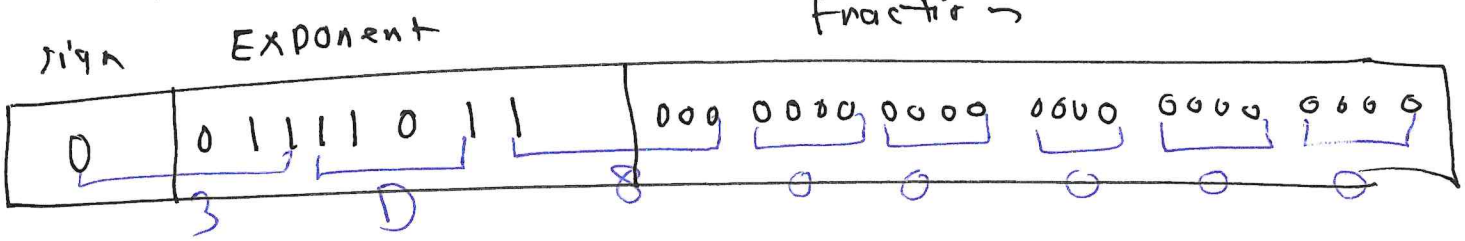
3. Normalized result $= 1.000 \times 2^{-4}$

IEEE-754 single precision format

$S = 0$

Exponent $= -4 + 127 = 123_{10} = 01111011_2$

Fraction $= 0.0$



Hex $= 0x3D800000$

Floating Point Multiplication

1. Add exponents:
2. Multiply significands:
3. Normalize result:
4. Round and renormalize if necessary:
5. Determine the sign of result from signs of operands

NB. We assumed that significands can be only 4 digits of the significands and two digits of the exponents .

Floating Point Multiplication (Decimal)

Consider a 4-digit decimal example: (1.110×10^{10}) X (9.2×10^{-5})

1. Add exponents

   e1 = 10 e2 = -5

   e = e1 + e2 = 10 − 5 = 5.

2. Multiply significands = 1.110 X 9.2

   = 10.212

   Result = 10.212 X 10^5.

3. Normalize result = 1.0212 X 10^6.

Solution

$0.5_{10} = 0.1$

Normalized $= 1.000 \times 2^{-1}$ ← e1

$0.4375_{10} = 0.0111$

Normalized $= 1.110 \times 2^{-2}$ ← e2

1. Add exponent)

$e1 = -1$ $\quad e2 = -2$

$e = e1 + e2 = -3$ [resulting exponent]

2. Multiply significand)

$1.000 \times 1.110 = 1.110$

result $= 1.11 \times 2^{-3}$

```
  1.000
  1.110
 -------
  0000
 1000
 1000
 1000
 -------
 1110000
```

3. Normalize result $= 1.11 \times 2^{-3}$

Problem: Perform binary floating-point multiplication and convert the resulting values to IEEE-754 single precision format. Finally convert them to hexadecimal values.

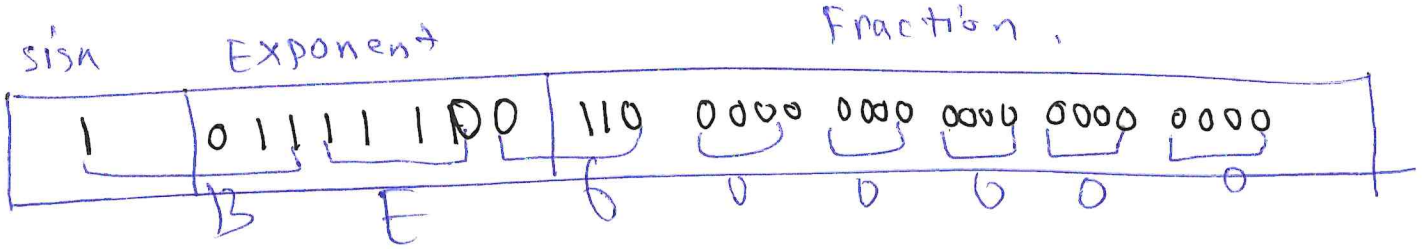
0.5 * -0.4375

4. sign = (+) * (−) = −

sign = 1

Exponent = $-3 + 127 = 124_{10} = 01111100$

Fraction = 0.11



0XBE600000