

**1,**

- (a) 123 = 01111011
- (b) 64 = 01000000
- (c) 127 = 01111111
- (d) 43 = 00101011
- (e) -123 = 10000101
- (f) -64 = 11000000
- (g) -127 = 10000001
- (h) -43 = 11010101

**2,**

- (a)  $123 - 127 = 01111011 - 01111111 = 01111011 + 10000001 = 11111100$
- (b)  $-123 - 127 = 10000101 - 01111111 = 10000101 + 10000001 = 00000110$  (overflow)
- (c)  $64 - 43 = 01000000 - 00101011 = 01000000 + 11010101 = 00010101$
- (d)  $43 - 43 = 00101011 - 00101011 = 00101011 + 11010101 = 00000000$
- (e)  $127 - 123 = 01111111 - 01111011 = 01111111 + 10000101 = 00000100$

**3,**

(a)  $9 = (1.001)_2 \times 2^3$

sign bit = 0

exponent code 01111111+00000011=10000010

mantissa code 0010000 00000000 00000000

IEEE754 single precision format: 41100000H

(b)  $5/32 = (1.01)_2 \times 2^{-3}$

sign bit = 0

exponent code 01111111+ 11111101 = 01111100

mantissa code 0100000 00000000 00000000

IEEE754 single precision format: 3E200000H

(c)  $-5/32 = -(1.01)_2 \times 2^{-3}$

sign bit = 1

exponent code 01111111+ 11111101 = 01111100

mantissa code 0100000 00000000 00000000

IEEE754 single precision format: BE200000H

(d)  $6.125 = (0110.001)_2 = (1.10001)_2 \times 2^2$

sign bit = 0

exponent code 01111111+00000010=10000001

mantissa code 1000100 00000000 00000000

IEEE 754 single precision format: 40C40000

**4,**

(a) 42E48000H = 01000010 11100100 10000000 00000000

sign bit = 0

exponent = 133 = 127 + 6

mantissa =  $(1.11001001)_2$  the value in decimal notation =  $(1.11001001)_2 \times 2^6 = 114.25$

(b) 3F880000H = 1.0625

(c) 00800000H = 0.0

(d) C7F00000H = -122880.0

**5,**

3EE00000H = 00111110 11100000 00000000 00000000

the exponent is 01111101

3D800000H = 00111101 10000000 00000000 00000000

the exponent is 01111011 = 01111101 - 00000010

the difference between the exponents is 2, align the significant of 3D800000H,

it becomes 00111110 10100000 00000000 00000000

the sum is 00111111 00000000 00000000 00000000 = 3F000000H

**6,**

(a) 125.25 - 75.5

$125.25 = (1.11110101)_2 \times 2^6$  to represent it in IEEE 754 format

sign bit = 0

exponent code 10000101

mantissa code 1111010 10000000 00000000

$75.5 = (1.0010111)_2 \times 2^6$  to represent it in IEEE 754 format

sign bit = 0

exponent code 10000101

mantissa code 0010111 00000000 00000000

their exponents are equal, the subtraction between mantissa is

$1.11110101 - 1.0010111 = 0.11000111 = (1.1000111)_2 \times 2^{-1}$

after normalization, the final result is 0 10000100 1000111 00000000 00000000 = 42470000H

(b)  $123.125 - 43.5$

$123.125 = (1.111011001)_2 \times 2^6$  to represent it in IEEE 754 format

sign bit=0

exponent code 10000101

mantissa code 1110110 01000000 00000000

$43.5 = (1.010111)_2 \times 2^5$  to represent it in IEEE 754 format

sign bit=0

exponent code 10000100

mantissa code 0101110 00000000 00000000

since their exponents are different, to operate subtraction between them, mantissa alignment is needed

$43.5 = (1.01011)_2 \times 2^5 = (0.1010111)_2 \times 2^6$

the mantissa subtraction  $1.111011001 - 0.1010111 = 1.001111101$

it is normalized format, the final result is 0 10000101 0011111 01000000 00000000  
= 429F4000H

7,

the real value here is  $Y=+0.4$

X will be the representation of Y in specifically defined floating format

Here exponent length is 4 bits, significant(mantissa) length is 7 bits.

We will represent Y by X as shown in the following:

$Y = +0.4 = +(0.0110011001100\dots)_2 = +(1.10011001100\dots)_2 \times 2^{-2}$

sign bit=0

exponent code 0111-0010 = 0101

significant code 1001100

for the 12-bit floating representation of Y is 0 0101 1001100

so, X = 0 0101 1001100 = 2CCH

the value of X according to such floating format is  $(1.1001100)_2 \times 2^{-2} = 0.3984375$

therefore, the relative error is expressed as

$R = (Y - X) / Y = (0.4 - 0.3984375) / 0.4 = 0.00390625$