HW1 - 20210774 김주은

1. Modified Problem of Exercise 2.61 on page 165.

- c파일에 코드 첨부

2. Exercise 2.73 on page 170

- c파일에 코드 첨부

3. Modified Problem of Exercise 2.77 on page 171

A. $23 = 2^4 + 2^3 - 2^0$

(x<<4) + (x<<3) – x

B. $-15 = 2^0 - 2^4$

x – (x<<4)

C. $19 = 2^4 + 2^1 + 2^0$

(x<<4) + (x<<1) + x

D. $-11 = 2^0 - 2^2 - 2^3$

X - (x<<2) – (x<<3)

4. Exercise 2.83 on page 172

A.

v = 0.yyyyy....

(v << k) = y.yyyy...

y.yyyy.... = v + $B2U_k(y)$ = v + Y

Y = (v << k) – v = v(2^k – 1) -> v = $\frac{Y}{2^\wedge k-1}$

B.

(a) 0.101 -> Y = 5, $\frac{5}{8-1} = \frac{5}{7}$

(b) 0.0110 -> Y= 6, $\frac{6}{16-1} = \frac{2}{5}$

(c) 0.010011 ->Y = 19, $\frac{19}{64-1} = \frac{19}{63}$

5. Exercise 2.82 on page 171

A. False, 반례 : x = INT_MIN, y=1 -> -x = INT_MIN 이므로 (x<y)!=(-x>-y)

B. True, (x+y)<<4 + y – x = (x+y)* 2^4 + y – x = 17*y + 15*x

C. True, x + ~x + 1 == 0 -> ~x + 1 == -x임을 이용하면

~(x+y) + 1 == -(x+y)

~(x+y) + 1 == -x -y

~(x+y) + 1 == ~x + 1 ~y + 1

-> ~(x+y) == ~x + ~y + 1

D. True, signed에서 unsigned로 cast할 때 bit representation은 변하지 않는다. 결국 x-y = -(y-x)와 같으므로 성립한다.

E. True, 2만큼 shift right하면서 LSB들이 버려지므로 다시 left shift 되었을 때 x값보다 작거나 같다.

6. Exercise 2.87 on page 173.

Description	Hex	M	E	V	D
-0	0X8000	0	-14	-0	-0.0
Smallest value > 2	0X4001	$\frac{1025}{1024}$	1	$\frac{1025}{512}$	2.00195312
512	0X6000	1	9	512	512.0
Largest denormalized	0X03FF	$\frac{1023}{1024}$	-14	$\frac{1023}{2^{24}}$	6.09755516e^-5
-∞	0XFC00	-	-	-∞	-∞
Number with hex Represention 3BB0	0X3BB0	$\frac{123}{64}$	-1	$\frac{123}{128}$	0.9609375

Sol)

1) -0 = 1 00000 0000000000

2) 2 = 1*2^1 -> E = 1, exp = 16 -> 0 10000 00...0

-> (smallest value > 2) = 0 10000 00...01 이므로, Hex = 4001 이고, M = 1.0000...01 = $\frac{2^{10}+1}{2^{10}} = \frac{1025}{1024}$

이를 통해 (smallest value > 2)의 값을 구하면 2*$\frac{2^{10}+1}{2^{10}} = \frac{2^{10}+1}{2^{9}} = \frac{1025}{512}$ 이다.

3) 512 = 2^9 -> sign bit = 0, M=1, E=9 -> exp = 9+15 = 24 , frac = 00...0 -> 0 11000 000...0

4) Largest denormalized = 0 00000 111...1 = 03FF -> exp = 0, E=-15, M=0.111....1=$\frac{\frac{1}{2}(1-\frac{1}{2^{10}})}{1-\frac{1}{2}} = \frac{(2^{10}-1)}{2^{10}} = \frac{1023}{1024}$

5) $-\infty$ = 1 11111 00...0 = FC00

6) 3BB0 = 0 01110 1110110000 -> exp = 14, E = 14 - 15 = -1, M = 1.1110110000 = $\frac{123}{64}$

7. Exercise 2.89 on page 174.

A. True. 똑같은 type 으로 casting 되므로 같은 precision 으로 변환되고, 값은 같다.

B. False, 예를 들어 x = 0, y = Tmin 이라면 (x-y)가 overflow 를 발생시키므로 dx-dy 와 (double)(x-y)의 결과가 다르다.

C. True, double 형은 commutative group 이기 때문에 결합법칙이 성립한다.

D. False, 결합법칙이 성립하지 않는다. (Tmax*(Tmax-2))*(Tmax-4) != Tmax* ((Tmax-2)*(Tmax-4)) 에서 각 좌변 우변에서 곱셈 후 근사과정에서 비트 손실과 오버플로우가 발생하면서 결과값이 달라진다.

E. False, dx == 0, dz !=0 인 경우 dx/dx 는 NaN 이므로 결과값이 같지 않다.

8. Exercise 2.91 on page 176

A. 0X40490FDB
-> 0100 0000 0100 1001 0000 1111 1101 1011 이므로
sign bit = 0, exp bit = 100 0000 0, frac bit = 100 1001 0000 1111 1101 1011 이다.
E = exp – bias = 2^7 – 127 = 1, M = 1. 10010010000111111011011 이다.
-> fractional binary number $=(11.0010010000111111011011)_2$

B.
$\frac{22}{7} = 2 * \frac{11}{7}$, $\frac{11}{7}$를 이진수로 표현하면 1.100100100100...
$\frac{22}{7} = 2 *\ 1.100100100100\ldots$
->fractional binary representation $=(11.001001001001001\ldots)_2$
C.
A에서의 floating point representation = $(11.0010010000111111011011)_2$
B에서의 floating point representation = $(11.001001001001001\ldots)_2$

소수점으로부터 9번째 bit position부터 diverge한다.

9. Exercise 2.92 on page 177

- c파일에 코드 첨부

10. Modified Problem of Exercise 2.95 on page 178

- c파일에 코드 첨부