① 0x222E4091

find sign and exp

0x222E4091

→ 0010 0010 0010 . . . .

s

\+

2^6 2^2

$2^6 + 2^2 - 127$

$= \boxed{-58}$

② Bin to 8-3 format
(no denormal)
bias: 3 total = 8
Exp = 3

$$\underbrace{0}_{+}\ \underbrace{001}_{e=1}\ \underbrace{0010}_{m}$$

$\rightarrow +1.0010 \times 2^{e-bias}$

$\rightarrow +1.0010 \times 2^{1-3}$

$\rightarrow +1.0010 \times 2^{-2}$

$\rightarrow 0.010010$

$\frac{1}{2}$ $\frac{1}{4}$. $\frac{1}{32}$

$\frac{1}{4} + \frac{1}{32} =$

③ 8 bin to 8-3 format (denormal)

bias : 3 total = 8

Exp = 3 M : 4

$\underbrace{0}_{S(+)}\ \underbrace{000}_{\text{denomal}}\ 0\ \underbrace{010}_{M}$

$\rightarrow$

if (denomal)

$\pm 0.M \times 2^{1-bias}$

$+0.0010 \times 2^{1-3}$

$\rightarrow 0.0010 \times 2^{-2}$

$\rightarrow 0.000010 \rightarrow \frac{1}{32} \rightarrow 0.03125$

$\frac{1}{2}$ $\frac{1}{4}$ $\frac{1}{32}$

④ Bin to float (Small)

total length = 12

Exp = 4

bias = 5 (denormal included)

m = 7

$\underset{+}{0}\ \underbrace{0000}_{\to\ d}\ 01001000$

$\Rightarrow$

if (d)

$\to +0.m \times 2^{1-\text{bias}}$

$\to +0.1001000 \times 2^{1-5}$

$\to +0.1001000 \times 2^{-4}$

$\to 0.00001001000$

$\rightarrow \frac{1}{32} + \frac{1}{256} >$

5

HW1.5. Float Ordering (8-3 format)

Consider a floating point format with

- 8 total bits
- 1 bit for the sign
- 3 bits for the exponent
- a bias of 3
- 4 bits for the mantissa
- *no* denormal encodings

Given the following floating point numbers in binary format, order them from **most negative** (top) to **most positive** (bottom).

(Try and see if you can do it without calculating the numbers.)

Drag from here:

− 10101100 E: 2 − bias = −1

− 11011101 E: 5 − bias = 2

+ 00011101

− 11001101 E: 4 − bias = 1

Construct your solution here:

11011101

11001101

10101100

00011101

Sign

if ($E_1 = E_2$) cmp M_1, M_2

⑥ Float to Bin

$F=8$ $\quad$ $bias=3$

$E=3$ $\quad$ denormal included

$M=4$

$$0.109375 = \frac{7}{64}$$

$$= \frac{1}{64} + \frac{2}{64} + \frac{4}{64}$$

$$= \frac{1}{64} + \frac{1}{32} + \frac{1}{16}$$

$$0.000.111$$

$$2^{-2}$$

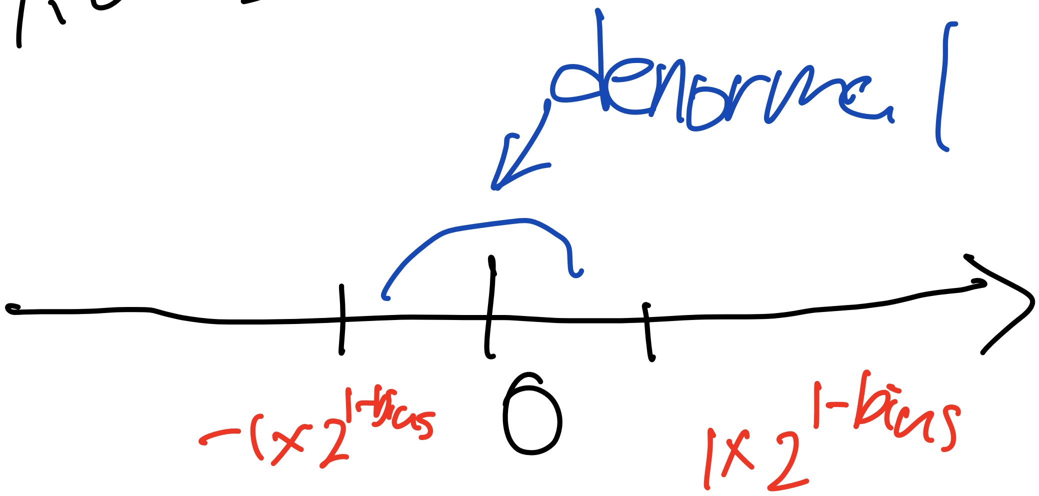
$$2^{x-bias} = 2^{-2}$$

$x=1$ $\quad$ 0.0111×2^{-2}

$$\underbrace{0}_{S}\ \underbrace{000}_{E}\ \underbrace{0111}_{M}$$

check if denormal

n: $\pm 1.0 \times 2^{1-bias}$



6.2

$-0.1875 = -\frac{3}{16}$

$= -\frac{1}{16} - \frac{2}{16}$

$= -\frac{1}{16} - \frac{1}{8}$

<u>denormal</u> ✓

-0.0011 $\quad 2^{1-\text{bias}}$

$= 2^{-2}$

$\longrightarrow -0.11 \times 2^{-2}$

floating format $\longrightarrow$

pitch 0

$\underbrace{1}_{(-)}\ \underbrace{0000}_{E}\ \underbrace{1100}_{M}$

⑦ Float to Bin
(no denormal)

$T = 11$
$E = 4$
$bias = 7$

$m = 6$

$$-0.34375 = -\frac{11}{32}$$

$$= -\left(\frac{1}{32} + \frac{2}{32} + \frac{8}{32}\right)$$

$$= -\left(\frac{1}{32} + \frac{1}{16} + \frac{1}{4}\right)$$

$\rightarrow -0.01011$

$\rightarrow -1.011 \times 2^{-2}$

$E - bias = -2$

$\rightarrow E - 7 = -2$

$\rightarrow E = 5$

$\rightarrow 0101$

$\underbrace{1}_{S}\ \underbrace{0101}_{E}\ \underbrace{011000}_{M=6}$

8 Float to Bin
(include denormals)

T = 9

E = 3

bias = 3

M = 5

$0.21875 = \frac{7}{32}$

$= \frac{1}{32} + \frac{2}{32} + \frac{4}{32}$

$= \frac{1}{32} + \frac{1}{16} + \frac{1}{8}$

$+0.00111$

$$2^{1-bias} = 2^{-2} = \frac{1}{4}$$

$$0.21875 < \frac{1}{4}$$

it's denormal

$$+0.00111$$

2^{1-bias} 形式 $\longrightarrow$ 0.111×2^{-2}

$$\underbrace{0}_{S}\ \underbrace{000}_{E}\ \underbrace{11100}_{M}$$

9) FP convert

format 1: $T=8$
$E=3$
$bias=3$
$m=4$

denormal included

→

format 2

$T=8$
$E=3$
$bias=7$
$m=4$

denormal included

Convert 0x02

↓

$$\underbrace{0}_{S}\,\underbrace{000}_{E}\,\underbrace{0010}_{M}$$

F1: $+0.0010\times 2^{1-3}$

$= 0.0010\times 2^{-2}$

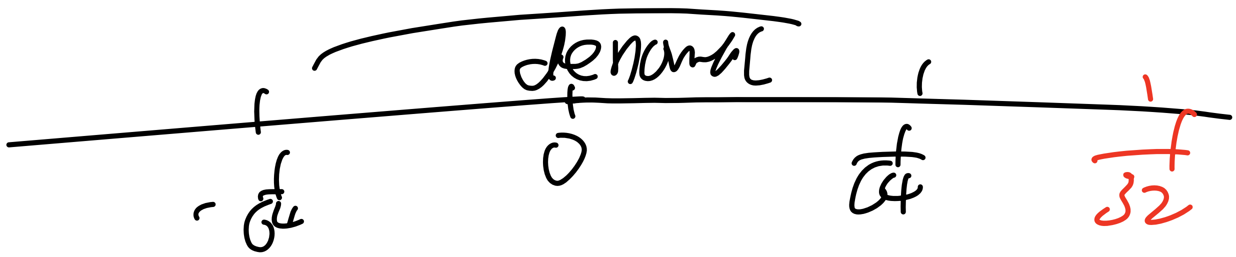
$= 0.000010$

$= \frac{1}{32}$

F2: denormal 范围

$2^{1-bias} = 2^{1-7} = 2^{-6}$

$\pm\frac{1}{64}$



not denormal

F2 (no denormal)

0.000010

$1.M \times 2^{x}$ 形式

1.0×2^{-5}

$E - 7 = -5$

$E = 2 \rightarrow 010$

$S \rightarrow 0$

$\underbrace{0}_{S}\ \underbrace{010}_{E}\ \underbrace{0000}_{M}$

$\downarrow$ hex

0x20

9.2

format 1:

$T=8$

$E=3$

bias $=4$

$m=4$

denormal included

f2:

$T=8$

$E=3$

bias $=6$

$m=4$

denormal included

f1

$$0X05 \rightarrow \underbrace{0}_{s}\underbrace{0000}_{E}\underbrace{101}_{m}$$

Since E : 000

it's denormal (f1)

$2^{1-bias} = 2^{1-4} = 2^{-3}$

$\rightarrow 0.M \times 2^{-3}$

$\rightarrow 0.0101 \times 2^{-3}$

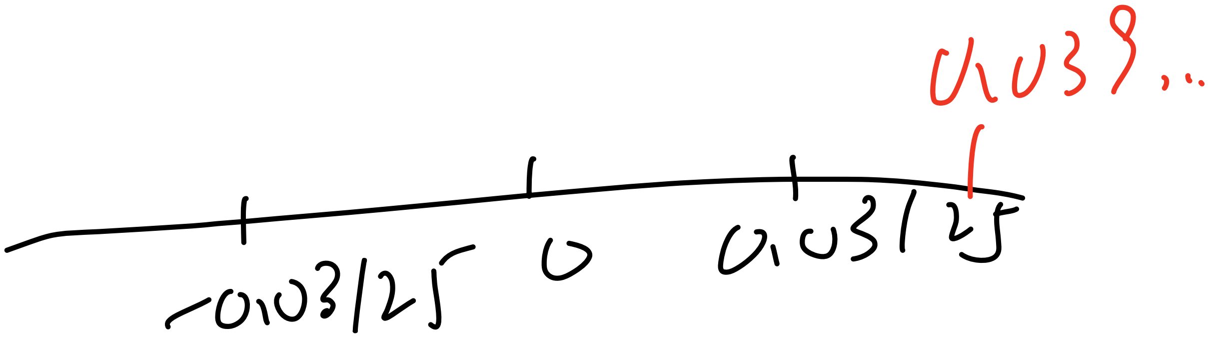
$\rightarrow 0.0000101$

32 128

$\rightarrow 0.0390625$

f2 denormal check

$2^{1-f2bias} = 2^{1-6} = 2^{-5}$



not denormal

→ 写成 $1.M \times 2^{E-bias}$

$1.\underbrace{01}_{m} \times 2^{-5}$

$E - 6 = -5$

$E = 1$

$E \to 001$

$S \to 0$

$\underbrace{0}_{s}\ \underbrace{001}_{E}\ \underbrace{0100}_{m} \xrightarrow{hex} 14$

10 FP Convert

f1: T=8
E=3
bias=3
m=4

denomal included

f2: T=10
E=3
bias=3
m=6

denormal included

$0x23 \rightarrow \underbrace{0}_{+}\ \underbrace{010}_{E}\ \underbrace{0011}_{m}$

not denormal

E = 2

$E-bias$

$\rightarrow 1.00111 \times 2$

$\rightarrow 1.00111 \times 2^{2-3}$

$\rightarrow 1.00111 \times 2^{-1}$

$\rightarrow 0.100111$

$\frac{1}{2} \quad \frac{1}{16} \quad \frac{1}{32}$

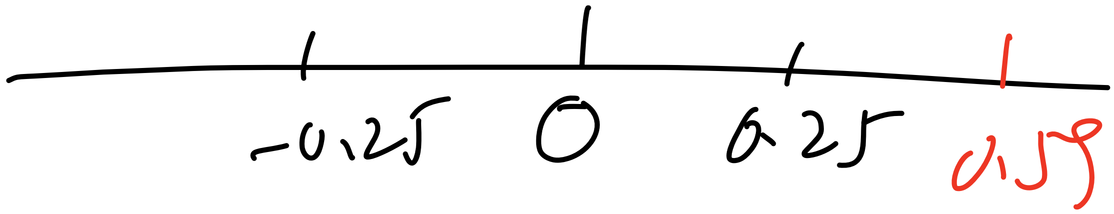
$\rightarrow \frac{1}{2} + \frac{1}{16} + \frac{1}{32}$

$\rightarrow 0.59375$

check format → denorm

$2^{1-\text{bias}} = 2^{1-3}$

$= 2^{-2}$



not denormal

0.10011

$1.m \times 2^{E-bias}$

1.0011×2^{-1}

$E - bias = -1$

$E - 3 = -1$

$E = 2$

(+)

0 010 001100

↓ hex

08C

(10).2

f_1: $T=8$

$E=3$

$bias=3$

$m=4$

include denormal

f_2: $T=12$

$E=4$

$bias-7$

$m=7$

include denormal

$0x8e \rightarrow \underbrace{1}_{S}\underbrace{000}_{E}\underbrace{1110}_{M}$

$\rightarrow$ denormal

f_1

$-0.1110 \times 2^{1-\text{bias}(f1)}$

$\rightarrow -0.1110 \times 2^{1-3}$

$\rightarrow -0.1110 \times 2^{-2}$

$\rightarrow -0.001110$

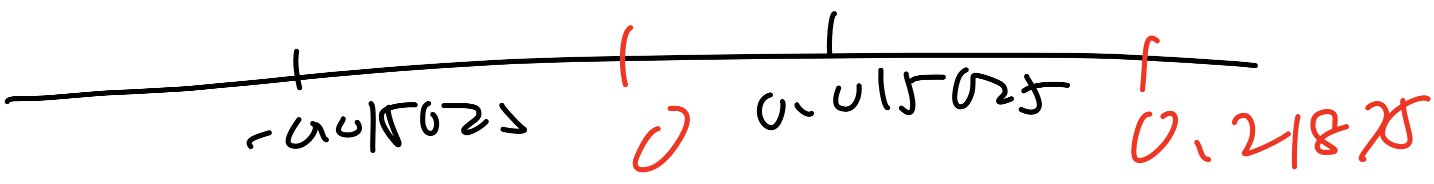
$\frac{1}{8}\ \frac{1}{16}\ \frac{1}{32}$

S.(-)

$\rightarrow 0.21875$

check f2 denormal

$2^{1-7} = 2^{-6} = 0.015025$



not denormal

-0.001110

$1.M \times 2^{E-bias} \rightarrow 1.\underbrace{110}_{M} \times 2^{-3}$

$E_2 - bias_2 = -3$

$E_2 - 7 = -3$

$E_2 = 4$

$\underline{(-)1}\ \underline{0100}\ \underline{110\,0000}$

hex

$\rightarrow$ A60

11

HW1.11. Format Comparison

Consider floating point formats Format-1 with

- 11 total bits
- 1 bit for the sign
- 4 bits for the exponent
- a bias of ***6***
- 6 bits for the mantissa
- *no* denormal encodings

and Format-2 with

- 12 total bits
- 1 bit for the sign
- 4 bits for the exponent
- a bias of ***4***
- 7 bits for the mantissa
- *no* denormal encodings

Which of the following formats can represent a *larger magnitude* number?

- (a) Format 1
- (b) Format 2
- (c) They can both represent the same largest magnitude number

Which of the following formats can represent a *smaller magnitude* number?

- (a) Format 1
- (b) Format 2
- (c) They can both represent the same smallest magnitude number

Suppose we changed both formats to support denormal representations. Would this change the answer to the previous (smaller magnitude) part? Pick the answer for this case.

- (a) Format 1
- (b) Format 2
- (c) They can both represent the same smallest magnitude number