HW03 2.1

The IEEE standard defines a binary representation for floating point using **sign**, **significant**, and **mantissa**.

Sign	Exponent	Significand
1 bit	8 bits	23 bits

For normalized floats:
Value = $(-1)^{Sign} \times 2^{(Exponent - Bias)} \times 1.significand_2$

For denormalized floats:
Value = $(-1)^{Sign} \times 2^{(Exponent - Bias + 1)} \times 0.significand_2$

Exponent	Significand	Meaning
0	Anything	Denorm
1-254	Anything	Normal
255	0	Infinity
255	Nonzero	NaN

$bias = (2^{(Exponent\ bits - 1)} - 1)$

$2^{8-1} - 1 = 2^7 - 1 = 128 - 1 = 127$

denorm Stride exp = 0

$2^{-126} \times 0.0\ldots1 = 2^{-126} \times 2^{-23} = 2^{-149}$ - smallest Pos denorm

$2^{-126} \times 0.0\ldots10 = 2^{-126} \times 2^{-22} = 2^{-148}$ - second Smallest Pos de norm

$2^{-148} - 2^{-149} = 2^{-148}(1 - 2^{-1}) = 2^{-148} \cdot 2^{-1} = 2^{-149}$ denorm Stride

Same

Smallest norm Stride : exp = 1

$2^{-126} \times 1.0\ldots0 = 2^{-126}$ - Smallest norm

$2^{-126} \times 1.0\ldots1 = 2^{-126} \times (1 + 2^{-23}) = 2^{-126} + 2^{-149}$ Smallest norm Stride.

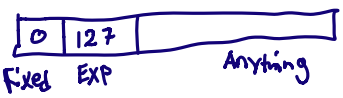
each time exponent increases, Stride doubles so when exp = 2 Stride $= 2^{-149} \cdot 2 = 2^{-148}$

Stride is usually $2^{exp - bias} \times 2^{-Significand\ Size}$

$2^{x-127} \times 2^{-23} = 1$

2^{x-150}

How many #'s between 1-2?



$2^{127-127} \times 1.0\ldots0 = 2^0 = 1$

$2^{127-127} \times 1.1\ldots1 = 2 - (2^0 \cdot 2^{-23}) = 2 - 2^{-23}$ ← Stride Size

next num $2^{128-127} \times 1.0\ldots0 = 2^1 = 2$

notice we have all values representable by Significand between $1-2 = 2^{23}$ things

Same between 2-4 there are 2^{23} #s. equally Spaced ruler between Powers of 2 with $2^{Sig\ Size}$ #'s between.

$1.0^{111\ldots11} = 6 - (2^2 \cdot 2^{-23}) = 6 - 2^{21}$

4-7

$2^2 = 2^{129-127} \times 1.000\ldots0 = 4$ (1/2, 23)

$2^3 = 8$?

$2^{25} \cdot 2^{-23}$

$2^2 \times 1.0111\ldots111$

$2^2 \times 1.11\,0\ldots0 = 7$ (21)

$(4-1) \cdot 2^{21} = 2^{23} - 2^{21} =$

$2^{23} - 2^{21}$

10 2^{21}

01 2^{21} ✓

Note:

$\sum_{i=0}^{n} 2^i = 2^{n+1} - 1$

$\sum_{i=0}^{23} 2^i = 2^{23} - 1$

2-4 $2^{23} - 2^{21} + 2^{22}$

00 [2^1]

11 [2^21] 7-8

2^{-1} 2^{-2}

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

3-7 2^{22}

$2^1 \times 1.1\underbrace{0\ldots0}_{22} = 3$

$2^1 \times 1.11\ldots1$

$2^2 \times 1.0\ldots0$

decimal ⇒ FP

39.5625

$1/2 = .5$

$1/4 = .25$

$1/8 = .125$

$1/16 = 0.0625$

$1/32 = 0.03125$

$1/64 = 0.015625$

① Convert to binary

$39 = 32+4+2+1 \rightarrow 100111.1001$

$0.5625 = 1/2 + 1/16$

② Convert to FP

$2^0 \times 100111.1001$

$= 2^5 \times 1.0011111001$ $\quad 132 = 128 + 4$

$2^5 = 2^{132-127}$

0b0 10000100 00111100 10...

132

$= 0X\ 42\ 1E\ 4000$

8.25 $\quad 8 + 1/4$

$2^0 \times 1000.01$

$2^3 \times 1.00001$

129

1	1000 0001	111 0...	0
S	Exp	Sig	

$(-1) \times 2^{129-127} \times 1.111$

$= -2^2 \times 1.111 = -111.1 = -7.5$

Example: Representing 1/3

- 1/3
  - $= 0.33333\ldots_{10}$
  - $= 0.25 + 0.0625 + 0.015625 + 0.00390625 + \ldots$
  - $= 1/4 + 1/16 + 1/64 + 1/256 + \ldots$
  - $= 2^{-2} + 2^{-4} + 2^{-6} + 2^{-8} + \ldots$
  - $= 0.0101010101\ldots_2 * 2^0$
  - $= 1.0101010101\ldots_2 * 2^{-2}$
  - **Sign**: 0
  - **Exponent** = -2 + 127 = 125 = 01111101
  - **Significand** = 0101010101…

0	0111 110	10101 0101 0101 0101 0101 010

Understanding the Significand (1/2)

- Method 1 (Fractions):
  - In decimal: 0.340_{10} $\Rightarrow 340_{10}/1000_{10}$ $\Rightarrow 34_{10}/100_{10}$
  - In binary: 0.110_2 $\Rightarrow 110_2/1000_2 = 6_{10}/8_{10}$ $\Rightarrow 11_2/100_2 = 3_{10}/4_{10}$
  - Advantage: less purely numerical, more thought oriented; this method usually helps people understand the meaning of the significand better

Understanding the Significand (2/2)

- Method 2 (Place Values):
  - Convert from scientific notation
  - In decimal:

$1.6732 = (1x10^0) + (6x10^{-1}) + (7x10^{-2}) + (3x10^{-3}) + (2x10^{-4})$

  - In binary:

$1.1001 = (1x2^0) + (1x2^{-1}) + (0x2^{-2}) + (0x2^{-3}) + (1x2^{-4})$

  - Interpretation of value in each position extends beyond the decimal/binary point
  - Advantage: good for quickly calculating significand value; use this method for translating FP numbers

Floating Point Fallacy

- FP add associative?
  - $x = -1.5 \times 10^{38}$, $y = 1.5 \times 10^{38}$, and $z = 1.0$
  - $x + (y + z) = -1.5\text{x}10^{38} + (1.5\text{x}10^{38} + 1.0)$
    $= -1.5\text{x}10^{38} + (1.5\text{x}10^{38}) = 0.0$
  - $(x + y) + z = (-1.5\text{x}10^{38} + 1.5\text{x}10^{38}) + 1.0$
    $= (0.0) + 1.0 = 1.0$
- Therefore, Floating Point add is not associative!
  - Why? FP result approximates real result!
  - This example: 1.5×10^{38} is so much larger than 1.0 that $1.5 \times 10^{38} + 1.0$ in floating point representation is still 1.5×10^{38}

Not associated

4.75

1/2 = .5
1/4 = .25
1/8 = .125
1/16 = 0.0625
1/32 = 0.03125
1/64 = 0.015625

100.11

$= 2^2 \times 1.0011 \quad 2^{1025-1023}$

s 0 q

EXP: 10000001

$2^{1023-1023} \qquad 2^{16}$

neg Pos

n	2^n
-1	0.5
-2	0.25
-3	0.125
-4	0.0625
-5	0.03125
-6	0.015625
-7	0.0078125
-8	0.00390625
-9	0.001953125
-10	0.0009765625
-11	0.00048828125
-12	0.000244140625
-13	0.0001220703125
-14	0.00006103515625
-15	0.000030517578125
-16	0.0000152587890625
-17	0.00000762939453125
-18	0.000003814697265625
-19	0.0000019073486328125
-20	0.00000095367431640625
-21	0.00000047683715820312 5
-22	0.00000023841857910156 25
-23	0.00000011920928955078125
-24	0.00000005960464477539062 5
-25	0.00000002980232387695312 5

$2^0 = 1$
$2^1 = 2$
$2^2 = 4$
$2^3 = 8$
$2^4 = 16$
$2^5 = 32$
$2^6 = 64$
$2^7 = 128$
$2^8 = 256$
$2^9 = 512$
$2^{10} = 1024$
$2^{11} = 2048$
$2^{12} = 4096$
$2^{13} = 8192$
$2^{14} = 16384$
$2^{15} = 32768$
$2^{16} = 65536$
$2^{17} = 131072$
$2^{18} = 262144$
$2^{19} = 524288$
$2^{20} = 1048576$
$2^{21} = 2097152$
$2^{22} = 4194304$
$2^{23} = 8388608$
$2^{24} = 16777216$
$2^{25} = 33554432$

4096-15 4080

3 exp bias = -3

4 mantisa. $2^{3-3} \times 1.0..0 = 2^0 = 1$

$2^0 \times 1.0...1 = 1(1+2^{-4})$

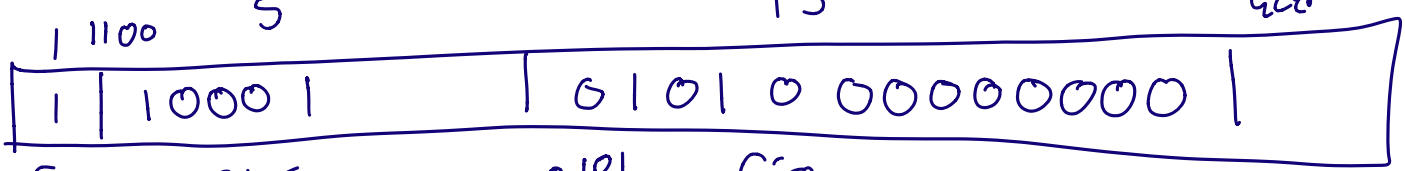
111

C
B
A

19 bits

1 1100 5

13



S exs 0101 Sig.

(-)

bias = -15

0X

-5.25

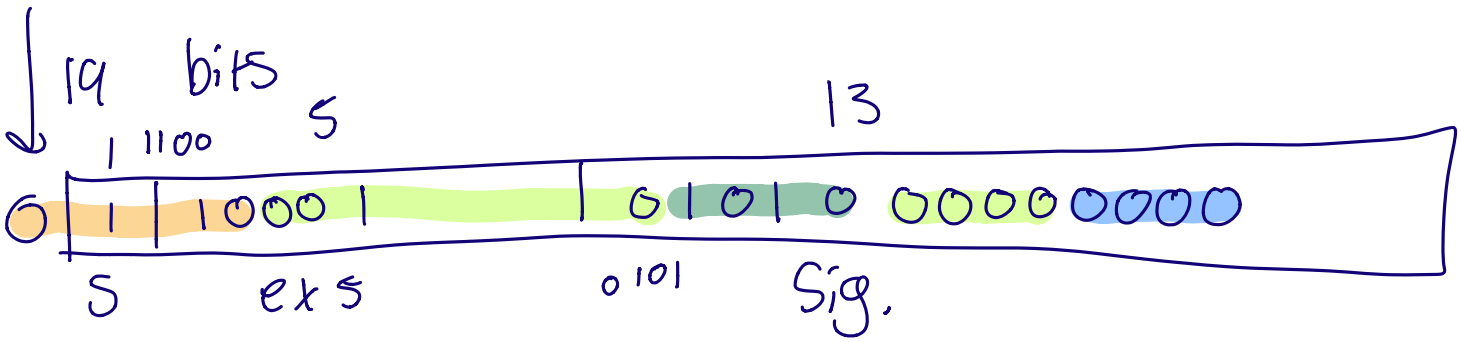
$2^0 \times 101.01$

$2^2 \times 1.0101$

0b 10001

2^{ex-15} ex = 17

Left Pad to convert



0110 0010 1010 0000 0000

0x 6 2 A 0 0

Part B

$6 = 4 \cdot 1.5$

$2^2 \times 1.0\ldots0 = 4$

12 → anything

$2^2 \times 1.10\ldots0 = 4(1+\frac{1}{2}) = 6$

$1.0\ldots1$

11

1010

decimal ⇒ FP 39.5625