Tutorial 4

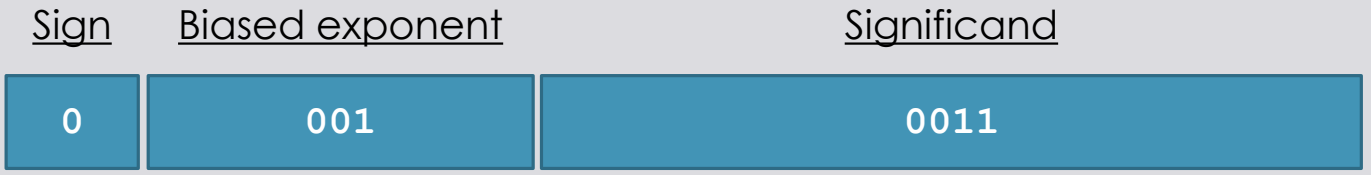
Floating point number representation and arithmetic

COMP2120B Computer organization

Kevin Lam (yklam2)

Floating point number representation

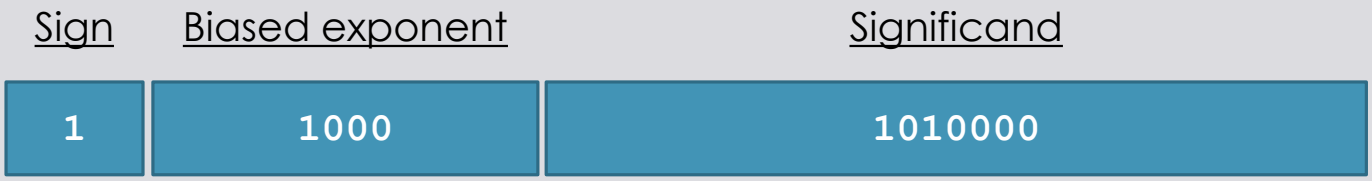
- Suppose we have an 8-bit floating point number with the following format:



- **Sign** (1 bit), **0** = positive, **1** = negative
  - In this case, it is a positive value
- **Biased exponent** ($k = 3$ bits), exponent = biased exponent – biased
  - biased = $2^{k-1} - 1 = 3$
  - exponent = $1 - 3 = -2$
- **Significand** ($N - k - 1 = 4$ bits), 1.*Significand*
  - $1.0011_2 = 1.1875$
- Therefore, the value above equals $1.1875 \times 2^{-2} = 0.296875$

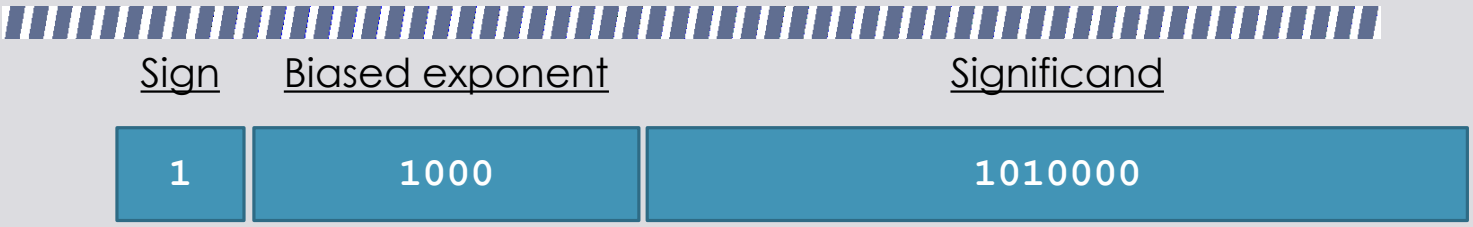
Exercise 1

a. Find the decimal value represented by this 12-bit binary pattern:



b. Find the corresponding binary pattern for the decimal value 10.125, using the above representation.

Exercise 1 - answer



a. $-1.101_2 \times 2^{8-7} = -1.625 \times 2 = -3.25$

Refer to the lecture slides on how to convert between decimal and binary

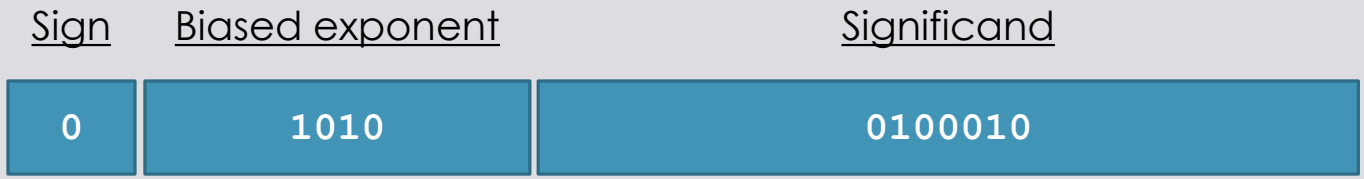
b. $10.125 = 1010.001_2 = 1.010001_2 \times 2^3 = 1.010001_2 \times 2^{10-7}$

Sign (positive): **0**

Biased exponent (decimal value 10): **1010**

Significand (1.010001): **0100010**

Remember to check the number of required bits



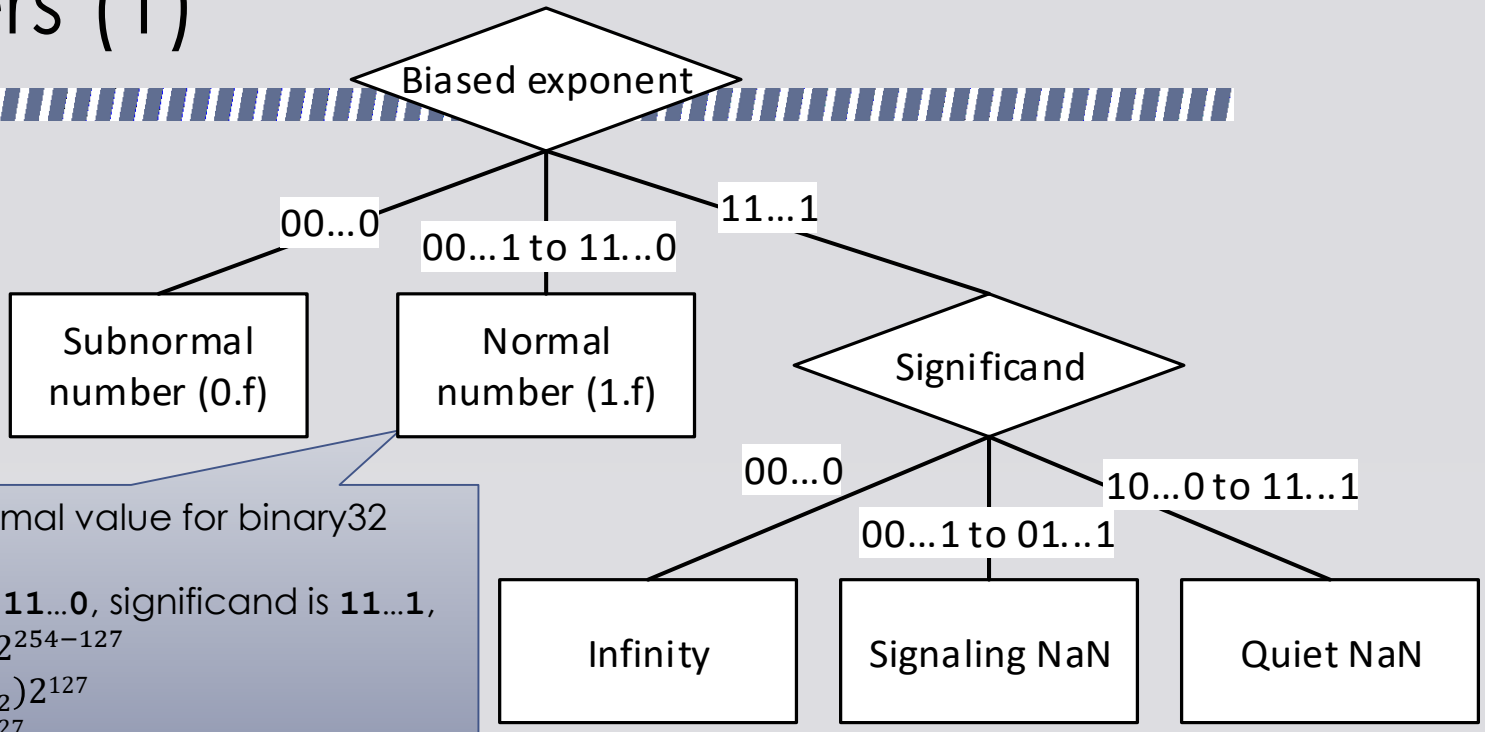
IEEE 754 basic binary formats

	Sign	Biased exponent	Significand
binary32	1	8	23
binary64	1	11	52
binary128	1	15	112

Number of bits

- What are the smallest and largest (non-zero) positive values that can be represented by the **binary32** format?
  - Bit pattern **`0 00000000 000…00`** and **`0 11111111 1111…11`** ?

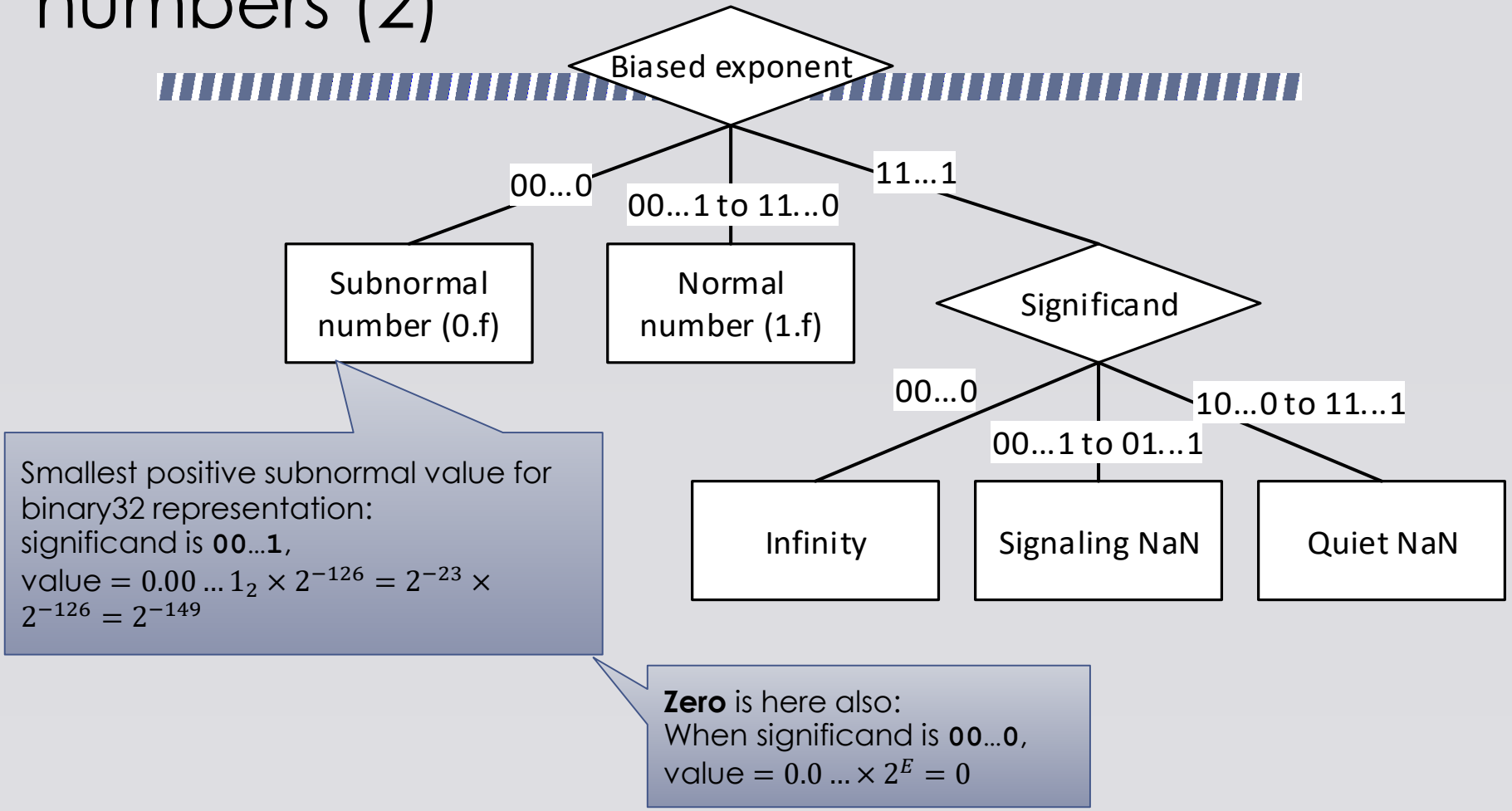
Interpretation of IEEE754 floating-point numbers (1)



Largest positive normal value for binary32 representation:
biased exponent is **11**…**0**, significand is **11**…**1**,
value $= 1.11\ldots1_2 \times 2^{254-127}$
$= (2 - 0.00\ldots1_2)2^{127}$
$= (2 - 2^{-23})2^{127}$
$= (1 - 2^{-24})2^{128}$

Smallest positive normal value for binary32 representation:
biased exponent is **00**…**1**,
significand is **00**…**0**,
value $= 1.0\ldots \times 2^{1-127} = 2^{-126}$

Interpretation of IEEE754 floating-point numbers (2)



Floating point multiplication/division
(the general flow)

- Multiplication:
  - Determine sign
  - Add biased exponent, subtract bias
  - Multiply significand
  - Normalize result
- Division:
  - Determine sign
  - Subtract biased exponent, add bias
  - Divide significand
  - Normalize result

Example	Sign	Biased exponent (bias = 3)	Significand
X (0.34375)	0 (+ve)	001 ($=1_{10}$)	0110 ($1.011_2=1.375_{10}$)
Y (-3)	1 (-ve)	100 ($=4_{10}$)	1000 ($1.1_2=1.5_{10}$)
$X \times Y$ (-1.03125)	1 (-ve)	Biased exponent $= 1 + 4 - 3 = 2_{10}$ Significand $= 1.375 \times 1.5 = 2.0625_{10} = 10.0001_2$	
Normalize result (shift right once)		Biased exponent $= 3$ Significand $= 1.00001_2$	
(-1 or -1.0625)	1	011 ($=3_{10}$)	0000 or 0001 (depending on rounding approaches)

Note that we assumed that we have enough guard bits for the calculation

Exercise 2

- Perform X/Y with the following representation.

	Sign	Biased exponent (bias = 7)	Significand
X (0.34375)	0 (+ve)	0101 ($=5_{10}$)	011 ($1.011_2=1.375_{10}$)
Y (-3)	1 (-ve)	1000 ($=8_{10}$)	100 ($1.1_2=1.5_{10}$)

	Sign	Biased exponent (bias = 3)	Significand
X (0.34375)	0 (+ve)	001 ($=1_{10}$)	0110 ($1.011_2=1.375_{10}$)
Y (-3)	1 (-ve)	100 ($=4_{10}$)	1000 ($1.1_2=1.5_{10}$)

Exercise 2 answer

Example	Sign	Biased exponent (bias = 7)	Significand
X (0.34375)	0 (+ve)	0101 ($=5_{10}$)	011 ($1.011_2=1.375_{10}$)
Y (-3)	1 (-ve)	1000 ($=8_{10}$)	100 ($1.1_2=1.5_{10}$)
X / Y (-0.114583…)	1 (-ve)	Biased exponent = $5-8+7 = 4_{10}$ Significand = $1.375/1.5 = 0.91666\ldots_{10} = 0.111010\ldots_2$	
Normalize result (shift left once)		Biased exponent = 3, Significand = $1.11010\ldots_2$	
(-0.109375 or -0.1171875)	1	0011 ($=3_{10}$)	110 or 111 (depending on rounding approaches)

Example	Sign	Biased exponent (bias = 3)	Significand
X (0.34375)	0 (+ve)	001 ($=1_{10}$)	0110 ($1.011_2=1.375_{10}$)
Y (-3)	1 (-ve)	100 ($=4_{10}$)	1000 ($1.1_2=1.5_{10}$)
X / Y (0.114583…)	1 (-ve)	Biased exponent = $1-4+3 = 0_{10}$ Significand = $1.375/1.5 = 0.91666\ldots_{10} = 0.111010\ldots_2$	
Normalize result (shift left once)		Biased exponent = -1, Significand = $1.11010\ldots_2$ ***Exponent underflow!***	

Floating point addition/subtraction
(the general flow)

- Addition:
  - Shift smaller exponent until both exponents equal
  - Add signed significands
  - Normalize result
- Subtraction (X-Y):
  - Change sign of Y
  - Do addition

Example	Sign	Biased exponent (bias = 3)	Significand	Signed significand
X (0.4375)	0	001 ($=1_{10}$)	1100	1.11_2 ($=1.75_{10}$)
Y (-3)	1	100 ($=4_{10}$)	1000	-1.1_2($=-1.5_{10}$)
X (shifting exponent)		100 ($=4_{10}$)		0.00111_2 ($=0.21875_{10}$)
$X + Y$ (-2.5625)		Biased exponent = 4, Significand = 0.00111_2 + (-1.1_2) = -1.01001_2 ($=-1.28125_{10}$)		
normalize result (do nothing)				
(-2.5 or -2.625)	1	100 ($=4_{10}$)	0100 or 0101 (depending on rounding approaches)	

Exercise 3

- Perform subtraction on the following

Example	Sign	Biased exponent (bias = 3)	Significand	Signed significand
X (28)	0	111 ($=7_{10}$)	1100	1.11_2 ($=1.75_{10}$)
Y (-12)	1	110 ($=6_{10}$)	1000	-1.1_2($=-1.5_{10}$)

Exercise 3 answer

Example	Sign	Biased exponent (bias = 3)	Significand	Signed significand
X (28)	0	111 ($=7_{10}$)	0110	1.11_2 ($=1.75_{10}$)
Y (-12)	1	110 ($=6_{10}$)	1000	-1.1_2($=-1.5_{10}$)
$-Y$ (12)	0	110 ($=6_{10}$)	1000	1.1_2($=1.5_{10}$)
$-Y$ (shifting exponent)		111 ($=7_{10}$)		0.11_2 ($=0.75_{10}$)
$X + (-Y)$ (40)		Biased exponent = 7, Significand = $1.11_2 + 0.11_2 = 10.1_2$ ($=2.5_{10}$)		
normalize result (shift right once)		Biased exponent = 8, Significand = 1.01_2 ($=1.25_{10}$) ***Exponent overflow!***		