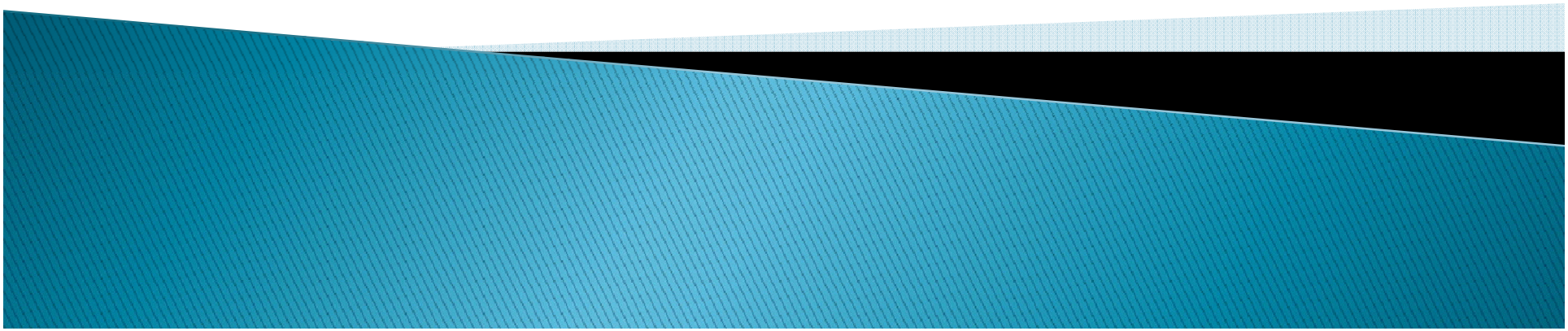
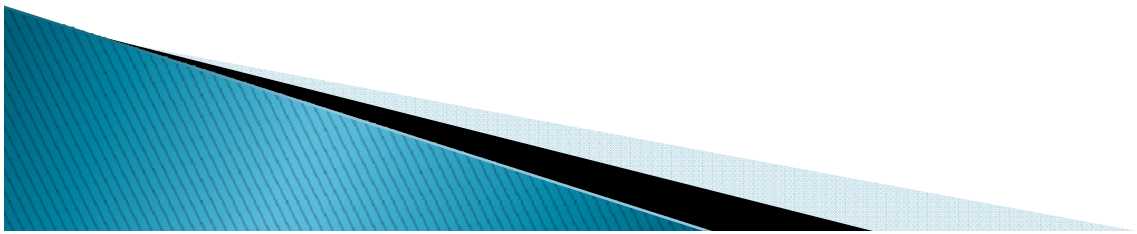
Compsci210 Tutorial

Data representation
IEEE 754 floating points

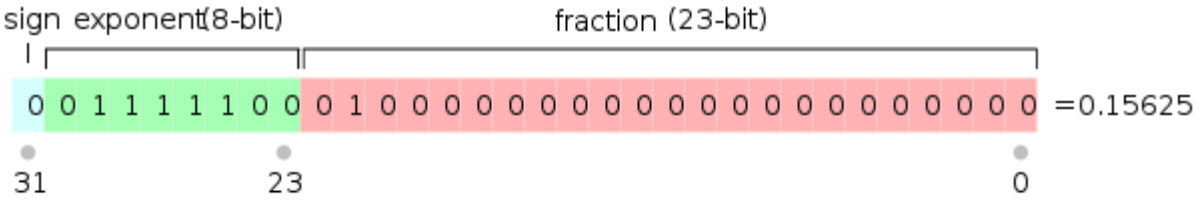


IEEE 754

- Generally, around 30% of exam will be on Data representation and the hardest parts and also most asked part of Data Representation will be on IEEE floating point number transformations and calculations.
- IE:
  - *Convert $C2100000_{16}$ from IEEE 754 Floating Point (Single Precision) to decimal*
  - *Convert 2.25 from Decimal to IEEE 754 Floating Point (Single Precision)*



IEEE 754 floating points structure



- 1 sign bit
- 8 exponent bits
- 23 mantissa bits
- Value of floating point number is in this form”
  - X = sign * (1.{mantissa}) * 2^{exponent – 127}

Part 1: convert from IEEE representation to Decimal float

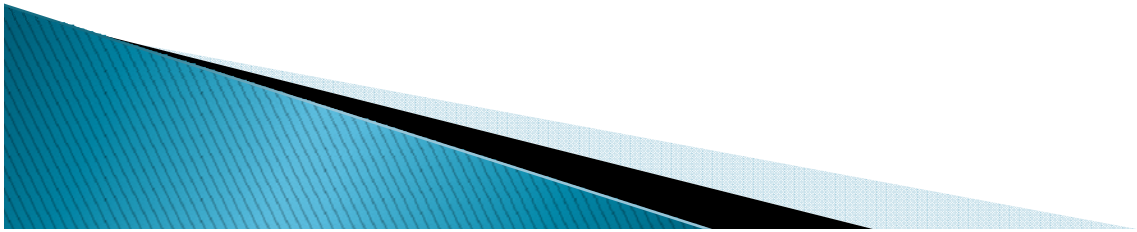
- *Convert $C2100000_{16}$*
  - Change this Hex to Bin
    - C2100000 = 1100 0010 0001 0000 0000 0000 0000 0000
  - Group this number in to 3 parts: sign, exp, man
    - 1100 0010 0001 0000 0000 0000 0000 0000
  - From this we can dig out these information:
    - Sign = 1 ➔ this number is a negative number
    - Exponent = 1000 0100 = 2^7 + 2^2 = 128+4 = 132
    - Mantissa = {00100…} = 1 + 2^-3 = 1 + 1/8 = 1.125
- Finally we got the answer:
  - X = -1 * 1.125 * 2^(132-127) = -1*1.125*2^5
    = -1*1.125 * 32 = -1 * 36 = -36.00

Part 2: convert from Decimal float to IEEE representation

- *8.625 from Decimal to IEEE 754 Floating Point*
  - *Change 8.625 to binary presentation:*
    - *8 = 1000_2*
    - *0.625 = 0.5 + 0.125 = 2^-1 + 2^-3 = 0.101_2*
    - *Hence 8.625 = 1000.101*
  - *Now we have to modify this number in term of*
    - *X = sign * {1}.{mantissa} * 2^(exponent127)*
    - *→ 1000.101 = 1.000101 * 2^3 //shift left by 3*
  - *Hence, we got:*
    - *Sign = 0 //positive number*
    - *Exp = 3+127 = 130 = 1000 0010*
    - *Mantissa = 0001 0100 0000 0000 …*
  - Finally group these 3 together:
    - X = 0100 0001 0000 1010 0000 0000 0…
    - X = **0x410A0000**

Part 3: IEEE-754 calculations

- Given X = 4130 0000, Y = 4050 0000, Evaluate X – Y in IEEE-754
- Step 1: Change X, Y to combinations of sign, exp and mantissa bits
  - X = 0100 0001 0011 0000 0000 0000 0000 0000
  - X = (+1) * 1.0110000 * 2^(100 0001 0)
  - Y = 0100 0000 0101 0000 0000 0000 0000 0000
  - Y = (+1) * 1.1010000 * 2^(100 0000 0)
- Step 2: Transform either X or Y so that both the number have the same exponent. Note exp of X = 10000010 and Y = 10000000, expX = expY + 2
  - X = (+1) * 1.0110000 * 2^(100 0001 0)
  - X = (+1) * 101.10000 * 2^(100 0000 0) //Move the dot 2 spaces to the right
- Step 3: Do calculation between 2 number:
  - X-Y = (+1) * 101.10000 * 2^(100 0000 0) – (+1) * 1.1010000 * 2^(100 0000 0)
  - X-Y = (+1) * 2^(100 0000 0) * (101.10000 – 1.1010000)
  - X-Y = (+1) * 2^(100 0000 0) * 11.111
  - X-Y = (+1) * 11.111 * 2^(100 0000 0) = (+1) * 1.1111 * 2^(100 0000 1)
- Step 4: Pick up the final values: sign bit, exp bits and mantissa bits
  - X-Y = 0 100 0000 1 1111 0000000000000
  - **X-Y = 4 0 F 8 0 0 0 0**



Exercises

Question 27

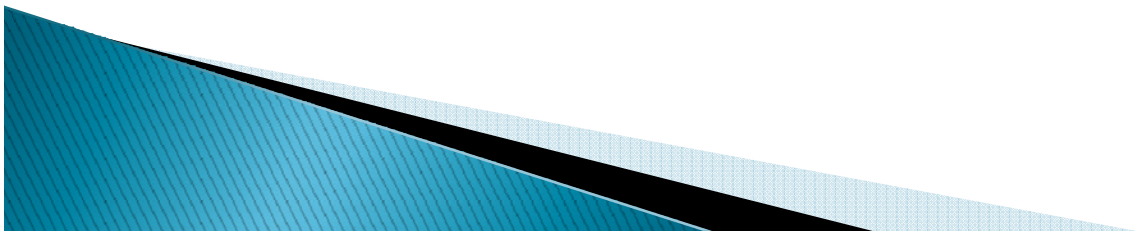
[1 mark]A 32-bit IEEE-754 floating-point number consists of 1 sign bit, 8 exponent bits and 23 mantissa bits. Given that `+0.1` is represented, in hexadecimal, as `3DCCCCCC`, give the first 12 binary digits of `-0.4`.

1. `1011 1101 1100`
2. `1011 1110 0100`
3. `1011 1110 1100`
4. `0011 1110 1100`

Question 28

[1 mark]A 32-bit IEEE-754 floating point number consists of 1 sign bit, 8 exponent bits and 23 mantissa bits. What decimal number is represented by `40D00000`?

1. `6.5`
2. `1.625`
3. `0.625`
4. `2.5`



Exercises

Question 37

[1 mark] A 32-bit IEEE-754 floating-point number consists of 1 sign bit, 8 exponent bits and 23 mantissa bits. Given that `FE400000` is represented IEEE floating point number in hexadecimal, what is the value of the exponent in base 2?

A. 126
B. 1.5
C. 252
D. 125
E. 124

Question 38

[3 marks] Given that A= `40A00000` and `B = 40E00000` are represented IEEE floating point numbers in hexadecimal. Evaluate `A + B`.

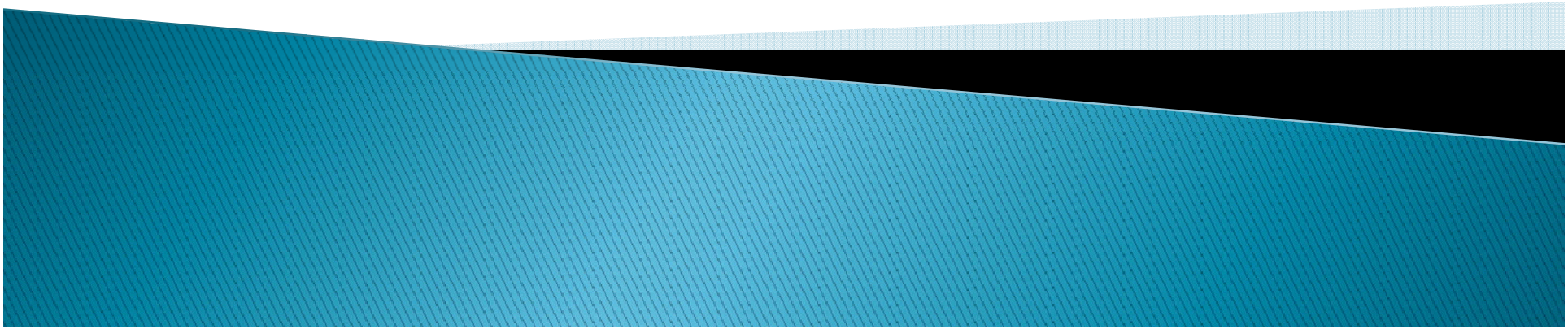
A. 41E00000
B. 41400000
C. C0400000
D. 40400000
E. 40C00000

Exercises

b) A 32-bit IEEE-754 floating-point number consists of 1 sign bit, 8 exponent bits and 23 mantissa bits. Given the following two IEEE-754 floating point numbers:
X = 41820000, Y = 3F200000, evaluate **X * Y**.
Express your result in hexadecimal. Show your working. (4 marks)

Compsci210 Tutorial IEEE754 exercise Answers

Minh Nguyen
Computer Science 210
2010 semester 1 City



Question 1

- +0.1 === 3DCCCCCC, find first 12 digits of -0.4
- 3DCCCCCC = 0011 1101 1100 …
- But we know
  - -0.4 is negative -> sign bit = 1
  - 0.4 = 4 * 0.1 = 2^2 * 0.1 -> exponent of 0.4 = 2^2 * exponent of 0.1
  - → exponent bits of 0.4
    = 011 1101 1_2 + 2_{10} = 011 1101 1_2 + 10_2
    = 01111101_2

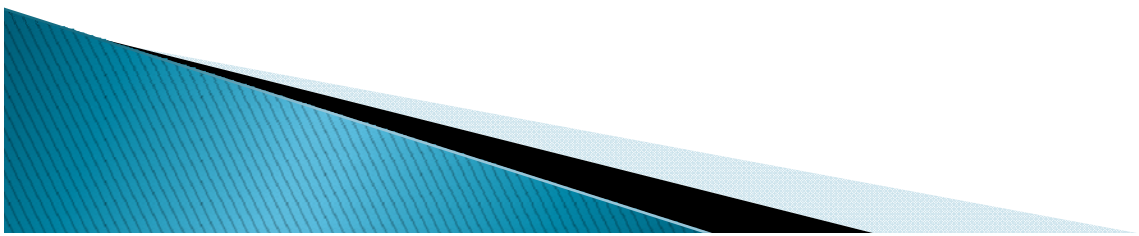
Group together == 1011 1110 1100

Question 2

- Turn 40D00000 to decimal
  - Convert to binary:
    - 0100 0000 1101 0000 0000 0000 0000 0000
  - Determine 3 components:
    - Sign bit = 0 -> positive number
    - Exponent bits = 100 0000 1
    - Mantissa bits = 10100…
  - Apply formula:
    - $= +1 * (1.101 * 2^{(100\ 0000\ 1 - 01111111)})$
    - $= 1.625 * 2^{129-127}$
    - $= 1.625 * 2^2$
    - $= 1.625 * 4$
    - $= 6.5$

Question3

- FE400000
- What is the value of the exponent in base 2?
- FE400000 = 1111 1110 0100 0000 …
  - Exponent bits:
  - = 111 1110 0
  - = 11111111 – 11
  - = 255 – 3 = 252
  - Remember this is biased by 127
  - → real exponent = 252–127 = **125**



Question 4

- A = 40A00000, B = 40E00000, find A+B
- 0100 0000 1010 0000 0000 0000 0000 0000
- 0100 0000 1110 0000 0000 0000 0000 0000
- Same exponent so do not need to normalise
- Doing addition on Mantissa:
  - 1.010 + 1.110 = 11.000
- → A+B
  - = $11.00 * 2^{100\ 0000\ 1}$
  - = $1.100 * 2^{100\ 0000\ 1+1}$
  - = $1.100 * 2^{100\ 0001\ 0}$
- Group → 0 10000010 1000000000…
- = 41400000

Question 5

- X = 41820000, Y = 3F200000, find X*Y
- 0100 0001 1000 0010 0000 0000 0000 0000
- 0011 1111 0010 0000 0000 0000 0000 0000
- Mantissa = man1*man2
- = 1.000 001*1.01
- = 1.01000101
- Exponent = exp1+exp2-bias
- = 100 0001 1 + 011 1111 0 - 01111111= 10000010
- Combine together
- X*Y = 0 10000010 01000101000000 =0x41228000