

- 1 (a) Provide the hexadecimal representation of a denormalized number in single precision IEEE 754 notation. What is the purpose of denormalized numbers?

A denormalized number is a floating point number that is smaller than the smallest number that can be represented in IEEE 754 notation. It is encoded by using a zero exponent value and a non-zero significand. An example of single precision denormalized number would be $0x00001100$. Recall that a zero exponent generally signifies the number 0.

- 1 (b) Assuming IEEE 754 rules and conventions, what is the result of the following operation in hexadecimal notation.

- $1.101 \times 2^{112} \times 1.01 \times 2^{12}$ (note that the exponent values are represented in base 10).

Remember that the exponents are biased. Thus when we add the exponents we subtract 127 to obtain the correct value of the biased exponent of the product. However when we do this we find that we have $112+12-127 = -3$. Since the value of the biased exponent is less than 0, we have underflow.

1 (c) Assuming that we have 6 significant digits (including the implicit bit), does the following calculation produce any error? If so what is the magnitude of the error?

$$1.001 \times 2^{121} + 1.1 \times 2^{128}$$

After adjusting the exponent value of the smaller number have

$$\begin{array}{r} 1.1000000000 \times 2^{128} + \\ 0.0000001001 \times 2^{128} \\ \hline 1.1000001001 \times 2^{128} \end{array}$$

With 6 significant digits the answer is 1.1×2^{128}

The magnitude of the error is 1.001×2^{121}

2 Perform the following floating point operations assuming that the number representations are IEEE 754 compliant. Your answers can be represented in scientific notation.

- *multiplication:* $0\text{xbe}000000 \times 0\text{xbf}800000$ (

Since the numbers are provided in hexadecimal notation we first convert them to decimal form.

$$-0.125 \times -1.0 = 0.125$$

- $1.001 \times 2^{131} + 1.11001 \times 2^{126}$

$$\begin{aligned} & 1.001 \times 2^{131} + 0.0000111001 \times 2^{131} \\ & = 1.0010111001 \times 2^{131} \end{aligned}$$

2 (a) Provide an example of a number that is too large to be represented in single precision notation but can be represented in double precision.

$$1.0 \times 2^{277}$$

3 (a) How would you interpret the following hexadecimal number assuming that it is a single precision IEEE 754 floating point number: 0x00140000?

It is a denormalized number since the exponent is 0, but the significand is non-zero.

3 (b) Assuming IEEE 754 rules and conventions, what is the result of the following operation in hexadecimal notation.

- $1.101 \times 2^{142} \times 1.01 \times 2^{122}$ (note that the exponent values are represented in base 10).

$$1.000001 \times 2^{138}$$

4 (a) Assuming a total of 4 significant digits including in the implicit bit, is there any inaccuracy in the following IEEE 754 computation? Justify your answer by showing how this value is calculated, and identifying the magnitude of the error if any.

$$(1.01 \times 2^{125}) \times (1.001 \times 2^{129})$$

$$1.001000 \times$$

$$1.01$$

$$1.01101 \times 2^{125+129-127=127}$$

Rounding down gives us (since the last two digits are 01)

$$1.011 \times 2^{127}$$

There is an error and the magnitude of the error is

$$0.00001 \times 2^{127}$$

4 (b) Assuming a total of 4 significant digits including in the implicit bit, is there any inaccuracy in the following IEEE 754 computation? Justify your answer by showing how this value is calculated, and identifying the magnitude of the error if any.

$$(1.01 \times 2^{125}) + (1.001 \times 2^{129})$$

$$0.000101 \times 2^{129}$$

$$1.001000 \times 2^{129}$$

$$1.001101 \times 2^{129}$$

Rounding up gives us (since the next two digits are 10)

$$1.01 \times 2^{129}$$

There is an error and the magnitude of the error is

$$0.000011 \times 2^{129}$$

5 (a) Assuming IEEE 754, what is the result of the following operations

- $1.1101 \times 2^{102} \times 1.01 \times 2^{172}$ (note that the exponent values are represented in base 10).

$$1.0010001 \times 2^{148}$$

5 (b) Consider a double precision floating point number (IEEE 754). As it turns out, the range of numbers that it can represent is not sufficient for your application. Keeping the total number of bits constant, how would you change the format of the number?

The number of bits in the exponent determines the range. This number should be increased at the expense of the number of bits in the significand.

5 (c) What is the smallest positive number you can represent in IEEE 754 double precision format. Represent this number in scientific notation.

$$1.0 \times 2^{-1022}$$

- 6 (a) Assuming a total of 6 significant digits including the implicit bit, is there any inaccuracy in the following computation? Justify your answer by showing how this value is calculated, and identifying the magnitude of the error if any.

$$(-1.11 \times 2^{-2}) + (1.01 \times 2^3)$$

The rounded value that is obtained is 1.0011×2^3 . The difference between this value and the actual value (assuming a sufficient number of bits) is given by 0.0000001×2^3 . The correct value is 1.0011001×2^3 .

- 6 (b) Assuming a total of 6 significant digits including the implicit bit, is there any inaccuracy in the following computation? Justify your answer by showing how this value is calculated, and identifying the magnitude of the error if any.

$$(-1.1 \times 2^{-1}) + (1.001 \times 2^4)$$

Correct answer: 1.000101×2^4

Rounded answer: 1.00011×2^4

Error: 0.000001×2^4

7 (a) Assuming IEEE 754, what is the result of the following operations

- $1.1101 \times 2^{82} \times 1.01 \times 2^{122}$ (note that the exponent values are represented in base 10).

$$1.0010001 \times 2^{78} \text{ (do not forget to subtract the extra 127 bias!)}$$

- $1.001 \times 2^{94} + 1.1 \times 2^{97}$ (note that the exponent values are represented in base 10).

$$1.101001 \times 2^{97}$$

7 (b) Assuming IEEE 754 rules and conventions, what is the result of the following operations. Assume 5 significant digits including the implicit bit.

- $1.11 \times 2^{12} \times 1.001 \times 2^{39}$.

$$2^{12} \times 2^{39} = 2^{51-127} = 2^{-76}$$

Remember that the exponent is biased and therefore we have to subtract 127 from the sum of the exponents. The fact that the result is negative indicates that the result is underflow.

- $1.0001 \times 2^{133} + 1.0101 \times 2^{139}$

$$1.0001 \times 2^{133} + 1.0101 \times 2^{139} =$$

$0.0000010001 \times 2^{139} + 1.0101 \times 2^{139} = 1.0101 \times 2^{139}$ (the result is rounded down according to the rules for 754).

7 (c) Provide an example of a number that is too small to be represented with single precision notation but can be represented with double precision.

1.0×2^{-130} This number is just smaller than the smallest number that can be represented in single precision numbers.

7 (d) What does the following bit pattern represent

00111100 00001001 01000000 00000000. Provide your answer in the form indicated.

A single precision IEEE 754 floating point number
(show actual value in scientific notation) _____ $1.000100101 \times 2^{-7}$ _____

7 (e) Assuming a total of 4 significant digits including in the implicit bit, is there any inaccuracy in the following IEEE 754 computation? Justify your answer by showing how this value is calculated, and identifying the magnitude of the error if any.

$$(1.11 \times 2^{114}) \times (1.01 \times 2^{119})$$

The product is 1.00011×2^{107}

With only 4 significant digits and the next two bits being 11, we round up to give

$$1.001 \times 2^{107}$$

The magnitude of the error is 0.00001×2^{107}