Quiz 2- Floating Points

Top of Form

**Timothy Beckett (username: tdb2q)**

**Attempt 2**

Written: Jun 4, 2023 3:15 PM - Jun 4, 2023 3:23 PM

**Submission View**

Released: Jun 6, 2023 7:00 AM

|  |  |  |
| --- | --- | --- |
| **Question 1** |  | 8 / 8 points |

Consider a 9-bit variant of the IEEE 754 floating point format as follows:

* Sign bit
* 4-bit exponent with a bias = 7
* 4-bit significand

All of the rules of IEEE 754 standard apply.

Fill in the numeric values represented by the following bit patterns. You **must**write your number in decimal form (e.g. 0.0146485375,-0.0146485375)

**010100111:**

\_\_\_11.5\_\_\_(25 %)

**100100100:**

\_\_\_-0.0390625\_\_\_(25 %)

**100101100:**

\_\_\_-0.0546875\_\_\_(25 %)

**000111000:**

\_\_\_0.09375\_\_\_(25 %)

|  |  |  |
| --- | --- | --- |
| **Question 2** |  | 2 / 2 points |

Which of the following 8 bit floating point numbers (1 sign, 3 exponent, 4 fraction) represent NaN?

|  |  |
| --- | --- |
|  | 1 000 1111 |
|  | 0 111 1111 |
|  | 0 100 0000 |
|  | 1 111 0000 |
| **Question 3** | |  | 2 / 2 points |

Assuming the following code:

int x = myfunction1(); /\* arbitrary value \*/

int y = myfunction2(); /\* arbitrary value \*/

int z = myfunction3(); /\* arbitrary value \*/

//convert to double

double dx = (double) x;

double dy = (double) y;

double dz = (double) z;

unsigned ux = x;

unsigned uy = y;

The following statement always evaluates to 1 (True or False):

dx - dy == (double) (x-y)

|  |  |  |
| --- | --- | --- |
|  |  | True |
|  |  | False |
| Hide question 3 feedback | | |
|  | |  |
| Lex x = 0 and y = Tmin  dx - dy will evaluates to +ve number  where as (x-y) will evaluates to negative number (tmin) | | |

|  |  |  |
| --- | --- | --- |
| **Question 4** |  | 2 / 2 points |

Assuming the following code:

int x = myfunction1(); /\* arbitrary value \*/

int y = myfunction2(); /\* arbitrary value \*/

int z = myfunction3(); /\* arbitrary value \*/

//convert to double

double dx = (double) x;

double dy = (double) y;

double dz = (double) z;

unsigned ux = x;

unsigned uy = y;

The following statement always evaluates to 1 (True or False):

(dx + dy) + dz ==  dx + (dy + dz)

|  |  |  |
| --- | --- | --- |
|  |  | True |
|  |  | False |
| Hide question 4 feedback | | |
|  | |  |
| Since each value ranges between Tmin and Tmax, their sum can be represented exactly. | | |

|  |  |  |
| --- | --- | --- |
| **Question 5** |  | 2 / 2 points |

Assuming the following code:

int x = myfunction1(); /\* arbitrary value \*/

int y = myfunction2(); /\* arbitrary value \*/

int z = myfunction3(); /\* arbitrary value \*/

//convert to double

double dx = (double) x;

double dy = (double) y;

double dz = (double) z;

unsigned ux = x;

unsigned uy = y;

The following statement always evaluates to 1 (True or False):

(dx \* dy) \* dz ==  dx \* (dy \* dz)

|  |  |  |
| --- | --- | --- |
|  |  | True |
|  |  | False |
| Hide question 5 feedback | | |
|  | |  |
| try dx and dy to be very large numbers and dz to be a small one. | | |

|  |  |  |
| --- | --- | --- |
| **Question 6** |  | 1 / 1 point |

Floating point addition follows associative property.

|  |  |  |
| --- | --- | --- |
|  |  | True |
|  |  | False |
| Hide question 6 feedback | | |
|  | |  |
| Check lecture 4 slide: "Mathematical Properties of FP Add" | | |

|  |  |  |
| --- | --- | --- |
| **Question 7** |  | 1 / 1 point |

Floating point multiplication follows monotonic property except for infinities and NAN.

|  |  |  |
| --- | --- | --- |
|  |  | True |
|  |  | False |
| **Question 8** | | |  | 0 / 2 points |

Round the following binary numbers to nearest 1/8. Perform round down action in case of "Half way":

a) 1011.1111001

\_\_\_1011.001\_\_\_Incorrect Response**(1100.000, 1100)**

b) 11001.101100000

\_\_\_11001.001\_\_\_Incorrect Response**(11001.101)**

|  |  |
| --- | --- |
|  |  |
|  | |

|  |
| --- |
| **Attempt Score:**  18 / 20 |
| **Overall Grade (last attempt):**  19 / 20 |

|  |
| --- |
|  |

Done

|  |
| --- |
|  |

Bottom of Form