Subject: Advanced Course On Computational Physics. Date:

Problem Set 2:

1: $5.5 + 10^{-8} = ?$

Representation of 5.5:

$5 \div 2 = 2$ r 1
$2 \div 2 = 1$ r 0
$1 \div 2 = 0$ r 1

$\Rightarrow 5_{10} = 101_2$

$0.5 \times 2 = 1.0$

$\Rightarrow 0.5_{10} = 0.1_2$

$\Rightarrow 5.5_{10} = 101.1_2$

Normalize the binary: $5.5_{10} = 1.011_2 \times 2^2$ → Exponent = 2+127 = 129

Single-precision floating representation of 5.5 $\quad 129_{10} = (10000001)_2$

0 10000001 01100000000000000000000

Representation of 10^{-8}

$\log_2(10^{-8}) \approx -26.575$

$\frac{10^{-8}}{2^{-27}} = 1.3479$

$\Rightarrow 10^{-8} \approx 1.3479 \times 2^{-27}$

$10^{-8}_{10} = (1.0101100 \times 2^{-27})_2$ → Exponent = −27+127 = 100

$1.3479_{10} = (1.01011001110101001011111)$

0 01100100 01011001110101001011111

aligning the exponents

the exponent differ by 129−100 = 29

shifting the mantissa of 10^{-8}

after shifting 29 bits to the right it becomes:

00000000000000000000101

29 > 23

Subject: Date:

after the mantissa of 10^{-8} is shifted and aligned with the exponent of 5.5 we can add the two mantissas

Mantissa of 5.5 : 01100000000000000000000

Mantissa of 10^{-8} after shifting

00000a....0

the result is simply the mantissa of 5.5 because the contribution from 10^{-8} is negligible due to the large diffrence in magnitude.

0110...-.0 + 0...-.0 = 0110....0

→ the final result of adding $5.5 + 10^{-8}$ in single precision floating-point format is simply **5.5**

2. A: $R = (1-2s)2^{e-Bias}(1.f)$ $1 \leq e \leq 254$

$R_{max} = (1) \times 2^{254-127} \times (1+2^{-23})$ Bias = 127

$R_{max} = (2-2^{-23}) \times 2^{127} \approx 3.40282 \times 10^{38}$

1 bit for sign
8 bit for exponent
23 bit for mantissa

Minimum Normalized Number

$R_{min} = (1) \times 2^{1-127} \times (1) = 2^{-126}$ * this number is normalized

$R_{min} \approx 1.1754943 \times 10^{-38}$

Minimum Denormalized Number.

$R_{min} = (1) \times 2^{1-127} \times 2^{-23} = 2^{-149} \approx 1.401298 \times 10^{-45}$

Subject: Date:

B: Double Precision :

- 1 bit for the sign
- 11 bits for the exponent with a bias of 1023
- 52 bits for the mantissa

$$R_{max} = (1)(2^{2046-1023})(1+1-2^{-52}) = 2^{2046-1023}(2-2^{-52}) \approx$$

$$R_{max} \approx 1.7976931348623157 \times 10^{308}$$

Minimum Positive Normalized Number

$$R_{min} = 2^{-1022}(1) = 2.2250738585072 \times 10^{-308}$$

Minimum Positive Denormalized Number

$$R_{min} = 2^{-1022}(2^{-52}) = 2^{-1074} \approx 4.9 \times 10^{-324}$$