Mohseen Mukaddam: mmukadda
Vishnu Narang : vnnarang

3. (5 pts) What would be the IEEE 754 single precision floating point representation of n = -34543210.0123459876543
21x10^{12} ? For explanation, I want you to document the steps your perform, in this order:
(1) What is n in decimal
fixed point form (ddd.ddddd);
(2) What is n in binary fixed point form (bbb.bbbb), storing the first 25 bits following the binary point;
(3) What is the normalized binary number, written in the form 1.bbbbb...bbb Å~ 2e, storing 25 bits following
the binary point?
(4) What are the 23 mantissa bits, after the bits in bit positions -24, -25, ... are eliminated using the round to nearest, ties to even mode; exclude the 1. part;
(5) What is the biased exponent in decimal and in binary?
(6) Write the 32-bits of the number in the order: s e m; and
(7) Write the final answer as an 8-hexdigit number.

Solution:

(1) Converting number to decimal point form -> 34543210012345987654.321 (ddd.dddd) [Fixed point form]

(2) Converting *n* to binary fixed point form :
1110111110110001000100010011111010000101110011001011100101000111 0.01010
0.321 * 2 = 0.642 (0)
0.642 * 2 = 1.284 (1)
0.284 * 2 = 0.568 (0)
0.568 * 2 = 1.136 (1)
0.136 * 2 = 0.272 (0)

(3) Normalized Binary number = 1.1101_1111_0110_0010_0010_0010_0 x 2^{63} [1.bbbb form]

(4) Mantissa bits (23) 1101_1111_0110_0010_0010_00*1(23)***0(24)0(25)** these bits round down to => 1101_1111_0110_0010_0010_001
Note: 24th bit: 0 and 25th bit: 0, hence round to nearest, ties to even mode results in 23th bit: 1.

(5) biased exponent (e) = 127 + 68 = 195 (decimal)
= 11000011 (binary) [8bits]

(6) s e m format = 1 1100_0011 1101_1111_0110_0010_0010_001
Note: s: 1 because number is *negative*
(7) final format: 1110_0001_1110_1111_1011_0001_0001_0001 (binary)
HEX: 0xE1EF_B111

4. (5 pts) What decimal floating point number does this big-endian IEEE 754 single precision number represent: n = 0xF4E3_C2D1?
For explanation, I want you to document the steps you perform, in this order:
(1) What is n in binary;
(2) What is the value of the sign bit; What does this value signify about the final number;
(3) What are the binary and decimal values of the biased exponent;
(4) What is the binary value of the mantissa, with the 1. Part preceding the binary point?
(5) What is the decimal value of the unbiased exponent;
(6) What is the decimal value of
the mantissa, with the leading 1. part?
(7) What is the final decimal real number, written in the form [-]d.ddddddddd dddddd Å~ 10e where d represents a decimal digit 0-9 and there is an optional leading negative sign.
Write exactly 15 digits after the decimal point (even if they are 0's) and round the final 15th digit up or down as required based on the value of the 16th digit (16th digit < 5 round down; otherwise, round up).

Solution:
(1) n in binary => 1111_0100_1110_0011_1100_0010_1101_0001
(2) Converting n to s m e format, we get =>
1 1110_1001 1100_0111_1000_0101_1010_001
Sign bit: 1 (*negative number*)
(3) Biased exponent (e) = 1110_1001 (Binary) => 233 (Decimal)
(4) Matissa => 1.1100_0111_1000_0101_1010_001 (Binary)
(5) Unbiased exponent = 233 - 127 = 106 (decimal)
(6) Decimal value of matissa = (1 + 0.1100_0111_1000_0101_1010_001)
Fraction = $1 * 2^{-1} + 1 * 2^{-2} + 0 * 2^{-3} + 0 * 2^{-4} + 0 * 2^{-5} + ...$
Therefore, Mantissa = (1 + 0.779382824897766113281 25) =
1.77938282489776611328125
(7) Final decimal number = (-1) * (1.77938282489776611328125) * 2^{106}
= converting 2^{106} to 10^{y} , we get y = 31.909179540
NOTE : ((log 2 * 106) / log 10) => y = 31.909 (approx)
= Decimal number => -1.779382824897766 * 10^{32} (approx)

5. (5 pts) What would be the IEEE 754 double precision floating point representation of 1.8275091565308567123856738 95965827169405837361 × 10-15. For explanation, I want you to document the steps you perform, in this order:
(1) What is n in decimal fixed point form (ddd.ddddd);
(2) What is n in binary fixed point form (bbb.bbbb), storing the first 110 bits following the binary point);
(3) What is the normalized binary number, written in the form 1.bbbbb...bbb × 2e , storing 54 bits following the binary point)
(4) What are the 52 mantissa bits, after the bits in bit positions -53, -54, ... are eliminated using the round to nearest, ties to even mode; exclude the 1. part;
(5) What is the biased exponent in decimal and in binary?
(6) Write the 64-bits of the number in the order: s e m; and
(7) Write the final answer as a 16-hexdigit number.

Solution:

(1) N in decimal fixed point form =
0.0000000000000018275091565308567123856738 95965827169405837361

(2) N in binary fixed point form =
0.00000000000000000000000000000000000000000000000010000011101011
111001011111111001110100110001100111

Calculation:
$1.827 \times 10^{-15} * 2 = 3.655 \times 10^{-15}$ (0)
$3.655 \times 10^{-15} * 2 = 7.310 \times 10^{-15}$ (0)
$7.310 \times 10^{-15} * 2 = 1.462 \times 10^{-14}$ (0)
$1.462 \times 10^{-14} * 2 = 2.924 \times 10^{-14}$ (0)
$2.924 \times 10^{-14} * 2 = 5.848 \times 10^{-14}$ (0)

(3) Normalized binary number =
$1.00000111010111110010111111110011101001100011001110000 \times 2^{-49}$

(4) After eliminating the 53rd and 54th bits using round to nearest, ties to even mode, n = 1.00000111010111110010111111110011101001100011001110**0**
Explanation: 53rd and 54th bits => 00 hence 52nd bit: **0**

(5) biased exponent = 1023 - 49 = 974 (decimal) => 011_1100_1110 (binary)

(6) s e m format => (s)0 (e) 011_1100_1110 (m) 0000_0111_0101_1111_0010_1111_1111_1001_1101_0011_0001_1001_1100

(7) Final format => 0011_1100_1110_0000_0111_0101_1111_0010_1111_1111_1001_1101_0011_0001_1001_1100

0x3CE0_75F2_FF9D_319C (HEX)