

Solutions - Homework 2

(Due date: February 9th @ 5:30 pm)

Presentation and clarity are very important!

PROBLEM 1 (10 PTS)

- Complete the following table. Use the fewest number of bits in each case.

You MUST show your conversion procedure. **No procedure = zero points.**

REPRESENTATION			
Decimal	Sign-and-magnitude	1's complement	2's complement
-129	11000 0001	101111110	101111111
-46	1101110	1010001	1010010
154	010011010	010011010	010011010
-64	11000000	10111111	1000000
-10	1001010	10101	10110

PROBLEM 2 (16 PTS)

- a) Perform the following additions and subtractions of the following unsigned integers. Use the fewest number of bits n to represent both operators. Indicate every carry (or borrow) from c_0 to c_n (or b_0 to b_n). For the addition, determine whether there is an overflow. For the subtraction, determine whether we need to keep borrowing from a higher bit. (6 pts)

Example ($n=8$):

$$\begin{array}{r} \checkmark \quad 54 + 210 \\ \begin{array}{r} c_8=1 \\ c_7=1 \\ c_6=1 \\ c_5=1 \\ c_4=0 \\ c_3=1 \\ c_2=1 \\ c_1=0 \\ c_0=0 \end{array} \\ \begin{array}{r} 54 = 0x36 = 0\ 0\ 1\ 1\ 0\ 1\ 1\ 0 \\ 210 = 0xD2 = 1\ 1\ 0\ 1\ 0\ 0\ 1\ 0 \end{array} \\ \hline \text{Overflow!} \longrightarrow 1\ 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0 \end{array}$$

$$\begin{array}{r} \checkmark \quad 77 - 194 \\ \text{Borrow out!} \longrightarrow \begin{array}{r} b_8=0 \\ b_7=0 \\ b_6=0 \\ b_5=0 \\ b_4=0 \\ b_3=1 \\ b_2=0 \\ b_1=0 \\ b_0=0 \end{array} \\ \begin{array}{r} 77 = 0x4D = 0\ 1\ 0\ 0\ 1\ 1\ 0\ 1 \\ 194 = 0xC2 = 1\ 1\ 0\ 0\ 0\ 0\ 1\ 0 \end{array} \\ \hline 0\ 0\ 0\ 0\ 1\ 0\ 1\ 1 \end{array}$$

$$\begin{array}{r} \checkmark \quad 244 + 267 \\ \checkmark \quad 39 + 218 \end{array}$$

$$\begin{array}{r} \text{No Overflow} \quad \begin{array}{r} c_8=0 \\ c_7=0 \\ c_6=0 \\ c_5=0 \\ c_4=0 \\ c_3=0 \\ c_2=0 \\ c_1=0 \\ c_0=0 \end{array} \\ \begin{array}{r} 244 = 0x0F4 = 0\ 1\ 1\ 1\ 1\ 0\ 1\ 0 \\ 267 = 0x10B = 1\ 0\ 0\ 0\ 0\ 1\ 0\ 1 \end{array} \\ \hline 511 = 0x1FF = 1\ 1\ 1\ 1\ 1\ 1\ 1\ 1 \end{array}$$

$$\begin{array}{r} \begin{array}{r} c_8=1 \\ c_7=1 \\ c_6=1 \\ c_5=1 \\ c_4=1 \\ c_3=1 \\ c_2=1 \\ c_1=0 \\ c_0=0 \end{array} \\ \begin{array}{r} 39 = 0x27 = 0\ 0\ 1\ 0\ 0\ 1\ 1\ 1 \\ 218 = 0xDA = 1\ 1\ 0\ 1\ 1\ 0\ 1\ 0 \end{array} \\ \hline \text{Overflow!} \longrightarrow 1\ 0\ 0\ 0\ 0\ 0\ 0\ 0\ 1 \end{array}$$

$$\begin{array}{r} \checkmark \quad 251 - 126 \\ \checkmark \quad 169 - 201 \end{array}$$

$$\begin{array}{r} \text{No Borrow Out} \quad \begin{array}{r} b_8=0 \\ b_7=1 \\ b_6=1 \\ b_5=1 \\ b_4=1 \\ b_3=0 \\ b_2=0 \\ b_1=0 \\ b_0=0 \end{array} \\ \begin{array}{r} 251 = 0xFB = 1\ 1\ 1\ 1\ 1\ 0\ 1\ 1 \\ 126 = 0x7E = 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0 \end{array} \\ \hline 125 = 0x7D = 0\ 1\ 1\ 1\ 1\ 1\ 1\ 0 \end{array}$$

$$\begin{array}{r} \text{Borrow out!} \longrightarrow \begin{array}{r} b_8=1 \\ b_7=1 \\ b_6=0 \\ b_5=0 \\ b_4=0 \\ b_3=0 \\ b_2=0 \\ b_1=0 \\ b_0=0 \end{array} \\ \begin{array}{r} 169 = 0xA9 = 1\ 0\ 1\ 0\ 1\ 0\ 0\ 1 \\ 201 = 0xC9 = 1\ 1\ 0\ 0\ 1\ 0\ 0\ 1 \end{array} \\ \hline 0xE0 = 1\ 1\ 1\ 0\ 0\ 0\ 0\ 0 \end{array}$$

- b) Perform the following operations, where numbers are represented in 2's complement arithmetic: (10 pts)

$$\begin{array}{ll} \checkmark \quad -70 + 63 & \checkmark \quad 490 + 47 \\ \checkmark \quad -257 + 256 & \checkmark \quad -127 - 183 \end{array}$$

- For each case:

- Determine the minimum number of bits required to represent both summands. You might need to sign-extend one of the summands, since for proper summation, both summands must have the same number of bits.
- Perform the binary addition in 2's complement arithmetic. The result must have the same number of bits as the summands.
- Determine whether there is overflow by:
 - Using c_n, c_{n-1} (carries).
 - Performing the operation in the decimal system and checking whether the result is within the allowed range for n bits, where n is the minimum number of bits for the summands.
- If there is overflow and we want to avoid it, what is the minimum number of bits required to represent both the summands and the result?

$n = 8$ bits	$c_8 \oplus c_7 = 0$	No Overflow	-70	$c_8 = 0$	$c_7 = 0$	$c_6 = 1$	$c_5 = 1$	$c_4 = 1$	$c_3 = 1$	$c_2 = 1$	$c_1 = 0$	$c_0 = 0$	$+$
			$63 =$	0	0	1	1	1	1	1	1	1	$+$
			$-7 =$	1	1	1	1	1	1	0	0	1	$+$

$-70 + 63 = -1 \in [-2^7, 2^7-1] \rightarrow$ no overflow

$n = 10$	bits
$C_{10} \oplus C_9 = 1$	$C_{10} = 0$
Overflow!	$C_9 = 1$
	$C_8 = 1$
	$C_7 = 1$
	$C_6 = 1$
	$C_5 = 0$
	$C_4 = 1$
	$C_3 = 1$
	$C_2 = 1$
	$C_1 = 0$
	$C_0 = 0$
490	= 0 1 1 1 1 0 1 0 1 0 +
47	= 0 0 0 0 1 0 1 1 1 1
	<hr/>
	1 0 0 0 0 0 1 1 0 0 1

$490+47 = 537 \notin [-2^9, 2^9-1] \rightarrow \text{overflow!}$

To avoid overflow:

$$\begin{array}{r}
 c_{11} + c_{10} = 0 \\
 \text{Overflow!} \\
 \hline
 \begin{array}{r}
 c_{11} = 0 \\
 c_{10} = 0
 \end{array}
 \quad
 \begin{array}{ccccccccc}
 c_9 & = 1 & c_8 & = 1 & c_7 & = 1 & c_6 & = 1 & c_5 = 0 \\
 490 & = 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 + \\
 47 & = 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \underline{\quad} 1 & 1 & 1
 \end{array}
 \end{array}$$

$490 + 47 = 537 \in [-2^{10}, 2^{10}-1] \rightarrow$ no overflow

$-257+256 = -1 \in [-2^9, 2^9-1] \rightarrow$ no overflow

$c_9 \oplus c_8 = 1$	$c_9 = 1$	$c_8 = 0$	$c_7 = 0$	$c_6 = 0$	$c_5 = 0$	$c_4 = 0$	$c_3 = 0$	$c_2 = 0$	$c_1 = 1$	$c_0 = 0$
Overflow!										
$-127 =$	1	1	0	0	0	0	0	0	0	1
$-183 =$	1	0	1	0	0	1	0	0	1	
	0	1	1	0	0	1	0	1	0	

$-127-183 = -310 \notin [-2^8, 2^8-1] \rightarrow \text{overflow!}$

To avoid overflow:

$c_{10} \oplus c_9 = 0$	$c_{10} = 1$	$c_9 = 1$	$c_8 = 0$	$c_7 = 0$	$c_6 = 0$	$c_5 = 0$	$c_4 = 0$	$c_3 = 0$	$c_2 = 0$	$c_1 = 1$	$c_0 = 0$
No Overflow											
-255	=	1	1	1	0	0	0	0	0	0	1
-230	=	1	1	0	1	0	0	1	0	0	1
-310	=	1	0	1	1	0	0	1	0	1	0

$-127 - 183 = -310 \in [-2^9, 2^9-1] \rightarrow$ no overflow

PROBLEM 3 (27 PTS)

- a) Calculate the result of the additions and subtractions for the following signed fixed-point numbers. Use the minimum number of bits for both operands and result so that overflow is avoided. (6 pts.)

The diagram illustrates the addition of two floating-point numbers using IEEE 754 format. The numbers are:

- $0.1110 + 1.010111$
- $11.11101 - 0.001101$

The result is $1.0001 + 1.001001$, which is 101.01101 .

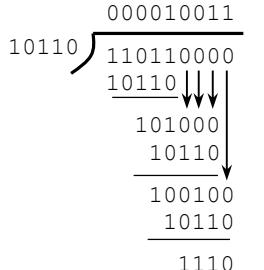
The diagram shows the step-by-step addition process, including alignment, addition of fractions, and normalization.

- b) Calculate the result of the multiplication of the following signed fixed-point numbers: (9 pts.)

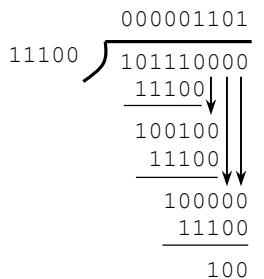
- c) Calculate the division result (with $x = 4$ fractional bits) for the following signed fixed-point numbers: (12 pts.)

$10.0101 \div 01.011$	$01.0111 \div 01.11$	$01.01110 \div 1.011$	$1.1101 \div 10.001$
-----------------------	----------------------	-----------------------	----------------------

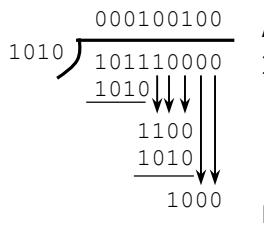
✓ $\frac{10.0101}{01.011}$: To positive (numerator), alignment, and then to unsigned: $a = 4: \frac{01.1011}{01.011} = \frac{01.1011}{01.0110} \equiv \frac{11011}{10110}$

Append $x = 4$ zeros: $\frac{110110000}{10110}$
 Unsigned Integer Division:

 $Q = 10011, R = 1110$
 $\rightarrow Qf = 1.0011 (x = 4)$ * Qf here is represented as an unsigned number
 Final result (2C): $\frac{10.0101}{01.011} = 2C(01.0011) = 10.1101$

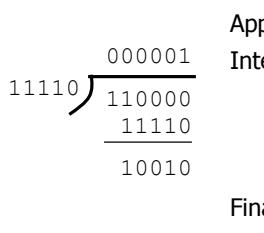
✓ $\frac{01.0111}{01.11}$: Alignment, and then to unsigned: $a = 4: \frac{01.0111}{01.1100} \equiv \frac{10111}{11100}$

Append $x = 4$ zeros: $\frac{101110000}{11100}$
 Unsigned Integer Division:

 $Q = 1101, R = 100$
 $\rightarrow Qf = 0.1101 (x = 4)$
 Final result (2C): $\frac{01.0111}{01.11} = 0.1101$ (this is represented as a signed number)

✓ $\frac{01.0111}{1.011}$: To positive (denominator), alignment, and then to unsigned, $a = 4: \frac{01.0111}{0.101} = \frac{01.0111}{0.1010} \equiv \frac{10111}{1010}$

Append $x = 4$ zeros: $\frac{101110000}{1010}$
 Integer Division:

 $Q = 100100, R = 1000$
 $\rightarrow Qf = 10.01 (x = 4)$ * Qf here is represented as an unsigned number
 Final result (2C): $\frac{01.0111}{1.011} = 2C(010.01) = 101.1100$

✓ $\frac{1.1101}{10.001}$: To unsigned and then alignment, $a = 4: \frac{0.0011}{01.111} = \frac{0.0011}{01.1110} \equiv \frac{11}{11110}$

Append $x = 4$ zeros: $\frac{110000}{11110}$
 Integer Division:

 $Q = 1, R = 10010$
 $\rightarrow Qf = 0.0001 (x = 4)$
 Final result (2C): $\frac{1.1101}{10.001} = 0.0001$

PROBLEM 4 (10 PTS)

- a) We want to represent numbers between -128.7 and 179 . What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0005 ? (3 pts.)

2C representation for integers: -2^{n-1} to $2^{n-1} - 1$. For $2^{n-1} - 1 \geq 179$, we have that $n \geq 9$, so we pick $n = 9$.

For the fractional part, we select the number of fractional bits p that make the resolution better or equal than 0.0005 :

$$2^{-p} \leq 0.0005 \rightarrow p \geq 10.0658 \rightarrow p = 11$$

Then the Fixed Point format required in [20 11].

- b) We want to represent numbers between -255.9 and 234.5 . What is the fixed point format that requires the fewest number of bits for a resolution better or equal than 0.0025 ? (3 pts.)

2C representation for integers: -2^{n-1} to $2^{n-1} - 1$. For $-2^{n-1} \leq -255.9$, we have that $n \geq 9$, so we pick $n = 9$.

For the fractional part, we select the number of fractional bits p that make the resolution better or equal than 0.0025 :

$$2^{-p} \leq 0.0025 \rightarrow p \geq 8.6439 \rightarrow p = 9$$

Then the Fixed Point format required in [18 9].

- c) Represent these numbers in Fixed Point Arithmetic (signed numbers). For each case, select the minimum number of bits

-127.3125	232.21875
-----------	-----------

- ✓ $127.3125 = 01111111.0101 \rightarrow -125.125 = 10000000.1011$
- ✓ $232.21875 = 011101000.00111$

PROBLEM 5 (9 PTS)

- a) Complete the table for the following fixed point formats (signed numbers):

Fractional bits	Integer Bits	FX Format	Range	Dynamic Range (dB)	Resolution
7	5	[12 7]	[-16, 15.9922]	66.23	0.0078125
12	4	[16 12]	[-8, 7.9998]	90.31	0.0002441
17	7	[24 17]	[-64, 63.99999]	138.47	0.00000763

- b) Complete the table for the following floating point formats (which resemble the IEEE-754 standard) with 16, 24, 48 bits. Only consider ordinary numbers.

$$\min = 2^{-2^{E-1}+2}, \max = (2 - 2^{-p})2^{2^{E-1}-1}, e \in [-2^{E-1} + 2, 2^{E-1} - 1], \text{significand} \in [1, 2 - 2^{-p}]$$

Exponent bits (E)	Significant bits (p)	Min	Max	Range of e	Range of significand
6	9	9.31×10^{-10}	4.29×10^9	$[-2^5 + 2, 2^5 - 1] = [-30, 31]$	$[1, 2 - 2^{-9}] = [1, 1.998046875]$
7	16	2.17×10^{-19}	1.84×10^{19}	$[-2^6 + 2, 2^6 - 1] = [-62, 63]$	$[1, 2 - 2^{-16}] = [1, 1.999984741210938]$
10	37	2.98×10^{-154}	1.34×10^{154}	$[-2^9 + 2, 2^9 - 1] = [-510, 511]$	$[1, 2 - 2^{-37}] = [1, 1.999999999992724]$

PROBLEM 6 (28 PTS)

- a) Calculate the decimal values of the following floating point numbers represented as hexadecimals. Show your procedure.

Single (32 bits)		Double (64 bits)	
✓ F8000378	✓ 7FFCDEAC	✓ 8009DECAE080000	✓ 7F00000000000000
✓ 801DECAF	✓ B300D959	✓ FFFDECAF0FEE90	✓ FACADEDECAE1990

✓ F8000378: 1111 1000 0000 0000 0000 0011 0111 1000

$$e + bias = 11110000 = 240 \rightarrow e = 240 - 127 = 113$$

$$\text{Mantissa } ([24 23]) = 1.0000000000001101111000 = 1.000105857849121$$

$$X = -1.000105857849121 \times 2^{113} = -1.0385693 \times 10^{34}$$

✓ 7FFCDEAC: 0111 1111 1111 1100 1101 1110 1010 1100

$$e + bias = 11111111 = 255, f \neq 0$$

$$X = NaN$$

✓ 801DECAF: 1000 0000 0001 1101 1110 1100 1010 1111

$$e + bias = 00000000 = 0 \rightarrow \text{Denormal number} \rightarrow e = -126$$

$$\text{Mantissa} = 0.00111011110110010101111 = 0.233785510063171$$

$$X = -0.233785510063171 \times 2^{-126} = -2.748135 \times 10^{-39}$$

b) Calculate the result of the following operations with 32-bit floating point numbers. Truncate the results when required. When doing fixed-point division, use 8 fractional bits. Show your procedure. (20 pts.)

✓ 40B00000 + C2FA8000 ✓ 10DAD000 - 90FAD000 ✓ 7AB80000 × 81800000 ✓ FA390000 ÷ 48400000
 ✓ 42FA8000 + COE00000 ✓ 3DE38866 - B300D959 ✓ FA09D300 × 4D080000 ✓ FF800000 ÷ 09FE0090

✓ $X = 40B00000 + C2FA8000:$
 $40B00000: 0100 \text{ } 0000 \text{ } 1011 \text{ } 0000 \text{ } 0000 \text{ } 0000 \text{ } 0000 \text{ } 0000$
 $e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2$ Mantissa = 1.011
 $40B00000 = 1.011 \times 2^2$

$$\begin{aligned} \text{C2FA8000: } & 1100\ 0010\ 1111\ 1010\ 1000\ 0000\ 0000\ 0000 \\ & e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6 \quad \text{Mantissa} = 1.11110101 \\ & \text{C2FA8000} = -1.11110101 \times 2^6 \end{aligned}$$

$$X = 1.011 \times 2^2 - 1.11110101 \times 2^6 = \frac{1.011}{2^4} \times 2^6 - 1.11110101 \times 2^6 = (0.0001011 - 1.11110101) \times 2^6$$

To subtract these unsigned numbers, we first convert to 2C:

$$\begin{array}{ccccccccc}
 C_1 & = & 0 & & & & & & \\
 C_9 & = & 0 & & & & & & \\
 C_8 & = & 0 & & & & & & \\
 & \vdots & & & & & & & \\
 C_6 & = & 0 & & & & & & \\
 C_5 & = & 1 & & & & & & \\
 C_4 & = & 1 & & & & & & \\
 C_3 & = & 1 & & & & & & \\
 C_2 & = & 1 & & & & & & \\
 C_1 & = & 0 & & & & & & \\
 C_0 & = & 0 & & & & & & \\
 \\[-1ex]
 \hline
 & 0 & 0.0 & 0 & 0 & 1 & 0 & 1 & 1 & 0 & + \\
 & 1 & 0.0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & \\
 & \hline
 & 1 & 0.0 & 0 & 1 & 0 & 0 & 0 & 0 & 1
 \end{array}$$

$$R = 0.0001011 - 01.11110101 = 0.0001011 + 10.00001011$$

The result in 2C is: $R = 10.00100001$, $-R = 01.11011111$

For floating point, we need the result in sign-and-magnitude:

$$\Rightarrow R(SM) = -1.11011111$$

$$X = -1.11011111 \times 2^6, e + bias = 6 + 127 = 133 = 10000101$$

$$X = \textcolor{blue}{1100\ 0010\ 1110\ 1111\ 1000\ 0000\ 0000\ 0000} = \text{C2EF800}$$

✓ $X = 42FA8000 + \text{COE}00000:$
 $42FA8000: 0100\ 0010\ 1111\ 1010\ 1000\ 0000\ 0000\ 0000$
 $e + bias = 10000101 = 133 \rightarrow e = 133 - 127 = 6$ Mantissa = 1.11110101
 $42FA8000 = 1.11110101 \times 2^6$

$$\text{COE}000000: 1 \ 100 \ 0000 \ 1110 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000 \ 0000$$

$$e + bias = 10000001 = 129 \rightarrow e = 129 - 127 = 2 \quad \text{Mantissa} = 1.11$$

$$\text{COE}000000 = -1.11 \times 2^2$$

$$X = 1.11110101 \times 2^6 - 1.11 \times 2^2 = 1.11110101 \times 2^6 - \frac{1.11}{2^4} \times 2^6 = (1.11110101 - 0.000111) \times 2^6$$

$$X = -1.000100111010011 \times 2^{117} \times 1.0001 \times 2^{27}$$

$$X = -1.0010010011100000011 \times 2^{144}$$

$$e + bias = 144 + 127 = 271 > 254$$

In this case, there is an overflow. The value X is assigned to $-\infty$.

$$X = 1\textcolor{red}{111} \ 1\textcolor{red}{111} \ 1\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} = \text{FF800000}$$

- ✓ $X = \text{FA390000} \div 48400000$:

$$\text{FA390000: } 1\textcolor{red}{111} \ 1\textcolor{red}{010} \ 0\textcolor{blue}{011} \ 1\textcolor{blue}{001} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000}$$

$$e + bias = 11110100 = 244 \rightarrow e = 244 - 127 = 117 \quad \text{Mantissa} = 1.0111$$

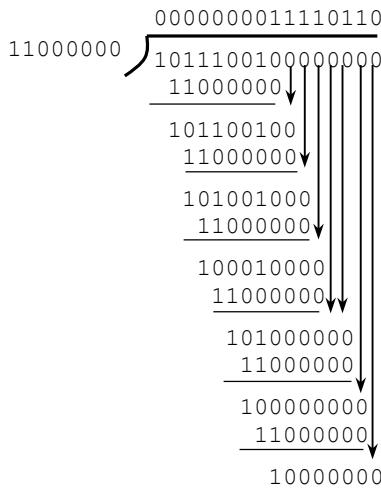
$$\text{FA390000} = -1.0111001 \times 2^{117}$$

$$48400000: 0\textcolor{red}{100} \ 1\textcolor{blue}{000} \ 0\textcolor{red}{100} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000}$$

$$e + bias = 10010000 = 144 \rightarrow e = 144 - 127 = 17 \quad \text{Mantissa} = 1.1$$

$$48400000 = 1.1 \times 2^{17}$$

$$X = -\frac{1.0111001 \times 2^{117}}{1.1 \times 2^{17}} = -\frac{1.0111001}{1.1} \times 2^{100}$$



Alignment:

$$\frac{1.0111001}{1.1} = \frac{1.0111001}{1.1000000} = \frac{10111001}{11000000}$$

Append $x = 8$ zeros: $\frac{1011100100000000}{11000000}$

Integer division

$$Q = 11110110 \rightarrow Qf = 0.1111011$$

$$\text{Thus: } X = -0.1111011 \times 2^{100} = -1.111011 \times 2^{99}$$

$$e + bias = 99 + 127 = 226 = 11100010$$

$$X = 1\textcolor{red}{111} \ 0\textcolor{red}{001} \ 0\textcolor{blue}{111} \ 0\textcolor{red}{110} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} = \text{F1760000}$$

- ✓ $X = \text{FF800000} \div 09FE0090$:

$$\text{FF800000: } 1\textcolor{red}{111} \ 1\textcolor{red}{111} \ 1\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000} \ 0\textcolor{blue}{000}$$

$$e + bias = 1111111 = 255, f = 0$$

$$\text{FF800000} = -\infty$$

$$X = -\infty \div \# = -\infty$$

$$X = \text{FF800000}$$