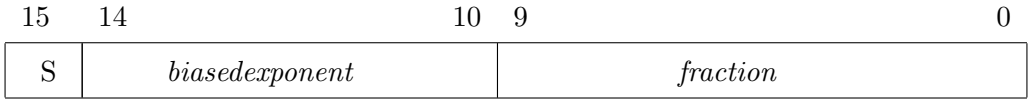
Question 3 (15 points):

A number in a 16-bit floating pointing format is represented as follows: the most significant bit is the sign bit, next there are 5 bits used for the exponent, and 10 bits for the fraction. This format is illustrated below:



The exponent is expressed in excess-16 format (also known as a bias representation). Given the binary representation above, the decimal value of the number represented can be computed by the following expression:

$$N = \begin{cases} (-1)^S \times 0.0 & \text{if } biasedexponent = 0 \text{ and } fraction = 0 \\ (-1)^S \times 0.fraction \times 2^{-14} & \text{if } biasedexponent = 0 \text{ and } fraction \neq 0 \\ (-1)^S \times 1.fraction \times 2^{biasedexponent-15} & \text{if } 0 < biasedexponent < 31 \\ (-1)^S \times \infty & \text{if } biasedexponent = 31 \text{ and } fraction = 0 \\ NaN & \text{if } biasedexponent = 31 \text{ and } fraction \neq 0 \end{cases}$$

1. (**10 points**) Complete the table below with the missing hexadecimal and decimal values for values in this representation.

Hexadecimal	Decimal
0xC808	-8.0625
0x0001	$2^{-24} = 5.96 \times 10^{-8}$
0x7C00	$+\infty$
0x7BFF	$2^{16}-2^5$

$8.0625 = 1000.0001 = 1.0000001 \times 2^3$
biasedexponent $- 15 = 3 \Rightarrow$ biasedexponent $= 18$
1 10010 0000001000 = 1100 1000 0000 1000 = 0xC808

0x0001 = 0000 0000 0000 0001 = 0 00000 0000000001 = $0.0000000001 \times 2^{-14} = 2^{-24}$

The printed version of the exam had a typo, the decimal value in the last row of the table was printed as $2^{31} - 220$. That number cannot be represented in the 16-bit floating point representation. The idea of the question was to have the following value in the last row:
$2^{16} - 2^5 = 0111111111100000 = 1.1111111111 \times 2^{15}$
biasedexponent $- 15 = 15 \Rightarrow$ biasedexponent $= 30$
0 11110 1111111111 = 0111 1011 1111 1111 = 0x7BFF

2. (**5 points**) In this representation `A = 0x7800` $= 32768_{10}$ and B = 0x4CC0 $= 19_{10}$ Assume a floating-point unit uses the NVIDIA format presented above. This unit has one guard, one round, and one sticky bit. What is the value of $A + B$, expressed in normalized base-two notation, computed by this machine?

`A = 0111 1000 0000 0000` $= 1.0 \times 2^{30-15} = 1.0 \times 2^{15}$
`B = 0100 1100 1100 0000` $= 1.0011 \times 2^{19-15} = 1.0011 \times 2^{4}$
Denormalizing B to make the exponents identical:
`B = 0.000 0000 0001 0011`

```
          mantissa       Guard  Round Sticky
  A = + 1.0000 0000 00|    0      0       0
  B = + 0.0000 0000 00|    1      0       1
-----------------------------------------------
A+B =   1.0000 0000 00|    1      0       1
```

Now we have to round up because of the sticky bit. Therefore the result is:
$A + B = 1.0000\ 0000\ 01 \times 2^{15} = 1000\ 0000\ 0010\ 0000_2 = 32768 + 32 = 32800_{10}$