Question 2 (30 points):

The 16-bit *half precision* floating point representation has the following specification:

15	14 10	9 0
S	*biasedexponent*	*fraction*

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

a. (**4 points**) what is the binary representation of -37.375 in the half-precision floating-point representation?

15	14 10	9 0
1	10100	0010 1011 00

$-37.375_{10} = -100101.011_2 = 1.00101011 \times 2^5$
$biasedexponent - 15 = 5 \Rightarrow biasedexponent = 20 = 01010_2$
$sign = 1$
$fraction = 0010101100$

Let `A = 0x7800` and B = 0x4D00 be two floating pointing numbers in this format.

b. (**8 points**) What is the value of `A` and the value of `B`? Express each of these values both in normalized base-two notation and in decimal notation.

	15	14 10	9 0
`A =`	0	11110	00 0000 0000

$A = (-1)^0 \times 1.0 \times 2^{30-15} = 1.0 \times 2^{15}$
$A = 1000\ 0000\ 0000\ 0000_2 = 2^{10} \times 2^5 = 1024 \times 32 = 32768_{10}$

	15	14 10	9 0
`B =`	0	10011	01 0000 0000

$B = (-1)^0 \times 1.01 \times 2^{19-15} = 1.01 \times 2^4$
$B = 10100_2 = 16 + 4 = 20_{10}$

c. **4 points**) What is the true value of $A + B$ expressed in decimal notation? In other words, what is the value of $A + B$ if an infinite precision could be used to compute the addition and to store the result?

$A + B = 32768 + 20 = 32788_{10}$

d. (**5 points**) Assume a floating-point unit uses the NVIDIA format presented above. This unit has no guard, no round, and no sticky bits. What is the value of $A + B$, expressed both in normalized base-two notation and in decimal notation, computed by this machine?

To align `A` with `B`, we need to move the binary point of `B` eleven positions to the left. Therefore: $B = 0.0000\ 0000\ 0010\ 1 \times 2^{15}$

```
          mantissa

  A = + 1.0000 0000 00

  B = + 0.0000 0000 00

----------------------------------------------

A+B =   1.0000 0000 00
```

Therefore $A + B = B = 1.0 \times 2^{15} = 32768_{10}$

e. (**5 points**) 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?

```
          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:

$A + B = 1.0000\ 0000\ 01 \times 2^{15} = 1000\ 0000\ 0010\ 0000_2 = 32768 + 32 = 32800_{10}$