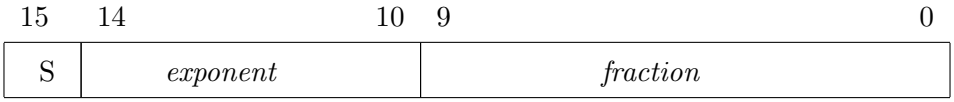
CMPUT 229 - Quiz # 5 - Fall 2011

Name: Solution

NVIDIA has defined a "half-precision" floating pointing format for use in its Graphics Processing Units (GPUs). A floating-point number is represented in this format in 16 bits 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:

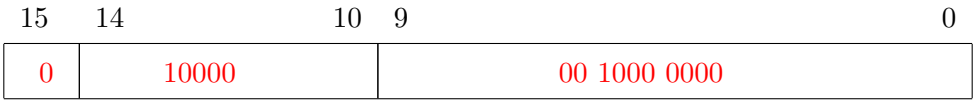


The exponent is expressed in excess-16 format (also known as a bias representation). Given the binary representation above, the decimal value of the number represented can be computed by the following expression:

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

1. (30 points) Give the bit pattern for the representation of the number 2.25_{10} in this notation.

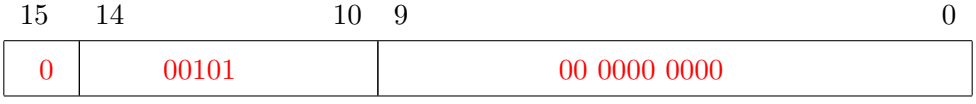
$2.25_{10} = 2.0 + 0.25 = 2 + \frac{1}{4} = (-1)^0 \times 10.01 = 1.001 \times 2^1$
$\Rightarrow exponent - 15 = 1 \Rightarrow exponent = 16$



Let `A = 0x000A` and B = 0x1400 be two floating pointing numbers in this format.

2. (20 points) What is the value, expressed in normalized base-two notation, of `A` and `B`?

$A = (-1)^0 \times 0.0000\ 0010\ 10 \times 2^{-14} = 1.01 \times 2^{-21}$



$B = (-1)^0 \times 1.0 \times 2^{5-15} = 1.0 \times 2^{-10}$

3. (25 points) Assume a floating-point unit with no guard, no round, and no sticky bits. What is the value of $A+B$, expressed in normalized base-two notation, computed by this machine?

   To align `A` with `B`, we need to move the binary point of `A` eleven positions to the left. Therefore:
   $A = 0.0000\ 0000\ 0010\ 1 \times 2^{-10}$

```
          mantissa
  A = + 0.0000 0000 00
  B = + 1.0000 0000 00
-----------------------------------------
A+B =   1.0000 0000 00
```

   Therefore $A + B = B = 1.0 \times 2^{-10}$

4. (25 points) Assume a floating-point unit with 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 = + 0.0000 0000 00|    1      0       1
  B = + 1.0000 0000 00|    0      0       0
-----------------------------------------------
A+B =   1.0000 0000 00|    1      0       1
```

   Now we have to round up because of the sticky bit. Therefore the result is:
   $A + B = 1.0000\ 0000\ 01 \times 2^{-10} \neq B$