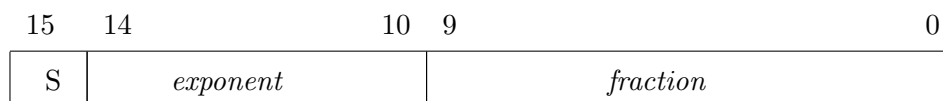


## CMPUT 229 - Quiz # 5 - Fall 2011

Name:

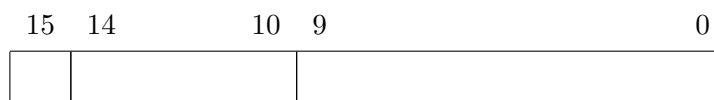
NVIDIA has defined a “half-precision” floating pointing format for use in its Graphics Processing Units (GPUs). A floating-point number is represented in this format in 16 bits as follows: the most significant bit is the sign bit, next there are 5 bits used for the exponent, and 10 bits for the fraction. 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:

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

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



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

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