CS 3843 Computer Organization Fall 2013

Solution to Quiz 2 Name (Last, First)_____, _____

Consider a hypothetical 8-bit floating point representation:

```
-----------------------------------------------
| s (1 bit) | exp ( 3 bits) | frac (4 bits)    |
-----------------------------------------------
```

(20 points) Questions:

1) (2 points) What is the bias? **Sol**: $2^{k-1} - 1 = 4-1 = 3$
2) (1.5 points) How many different values can be represented with 8 bits?
   Sol: $2^8=256$
3) (1.5 points) How many of these are NaN?
   Sol: From [0 111 0001] through [0 111 1111] there are 15 and
   From [1 111 0001] through [1 111 1111] there are 15 totally 30.
4) (1.5 points) How many of these are Infinity?
   Sol: 2. [0 111 0000] is +∞ and [1 111 0000] is -∞.
5) (1.5 points) How many of these are normalized values?
   Sol: Sign can take 2 different values. Exponent can take 6 different combinations. Fraction can take 16 combinations. Totally 2×6×16 = 192.
6) (1.5 points) How many of values are zero (denormalized)?
   Sol: 2. [0 000 0000] represents +0.0 and [1 000 0000] represents -0.0
7) (1.5 points) How many of these are non-zero denormalized values?
   Sol: 30. From [0 000 0001] to [0 0000 1111]. From [1 000 0001] to [1 000 1111].
8) (2 points) What is the smallest positive normalized value?
   Sol: $k = 3$, $n = 4$
   $s = 0$, $exp = 001$, $frac = 0000$;
   $M = 1+frac\times 2^{-n}=1$ $E = exp - Bias = 1 - 3 = -2$
   $V = (-1)^s \times M \times 2^E = 1\times 1\times 2^{-2} = \frac{1}{4}$
9) (2 points) What is the largest positive normalized value?
   Sol: $k = 3$, $n = 4$
   $s = 0$, $exp = 110$, $frac = 1111$;
   $M = 1+frac\times 2^{-n}=1+1111\times 2^{-4}=1.1111$ $E = exp - Bias = 6-3=3$

$V = (-1)^s \times M \times 2^E = 1 \times 1.1111 \times 2^3 = 1111.1 = 15.5$

10) (2 points) What is the largest positive denormalized value?

Sol: $k = 3$, $n = 4$

$s = 0$, $exp = 000$, $frac = 1111$;

$M = frac \times 2^{-n} = 1111 \times 2^{-4} = 0.1111$; $E = 1 - Bias = 1-3 = -2$

$V = (-1)^s \times M \times 2^E = 1 \times 0.1111 \times 2^{-2} = 1111 \times 2^{-6} = \frac{15}{64}$

11) (3 points) What is the floating-point representation for 1.0?

Sol: $1 = 1.0 \times 2^1 = M \times 2^E$

$M = 1.0 = 1 + frac \times 2^{-n}$ => $frac = 0000$

$exp = E + Bias = 0 + 3 = 3$ => $exp = 011$

So its floating-point representation is 0 011 0000