CS 3843 Computer Organization, Fall 2013 Assignment 3 Solution

Due Friday, October 4, 2013

This assignment is due at the beginning of class on the due date. There will be a 10 percent penalty for late assignments.

Solve the following problems and hand them in at the beginning of class on the due date. You may use a basic calculator, but you must show how you got your answer.

Write out the solutions neatly. Problems should be in order. Your **name** and the assignment number should be in the upper right corner of the first sheet you hand in. Stack the pages neatly and put a single stable in the upper left corner. Make sure the staple does not obscure any of your writing.

1. (40 points) Given a floating-point format with one sign bit, four exponent bits (k=4), and three fraction bits (n=3). Answer the following questions:
   a. (4 points) What is the bias?

      Sol: Bias = $2^{k-1}-1=2^{4-1}-1=2^3-1=7$

   b. (3 points) How many different values can be represented with 8 bits?

      Sol: $2^8=256$

   c. (3 points) How many of these values are NaN?

      Sol: s = 0 or 1, *exp* = 1111, *frac* ≠ 0, then the choices of *frac* are 001, 010, 011,100,101,110, 111, in total 7.

      Therefore, in total 2×7=14

   d. (3 points) How many of these are infinity?

      Sol: s = 0 or 1, *exp* = 1111, *frac* = 0000

      Therefore, in total 2 (±∞).

   e. (3 points) How many of these are positive, normalized value?

      Sol: s = 0, *exp* ≠ 0000, or 1111, the choices of *exp* are from 0001 to 1110, in total 2^4-2=14. The choices of *frac* is from 000 to 111, in total 2^3=8.

      Therefore, in total 14×8 = 112.

   f. (3 points) How many of these are negative, normalized value?

      Sol: similar as above, except s = 1.

      Therefore, in total 14×8 = 112.

   g. (3 points) How many of these are zero(denormalized)?

Sol: s = 0 or 1, exp = 0000, $frac$ = 0000

Therefore, in total 2.

h. (3 points) How many of these are positive, denormalized value?

Sol: s = 0, exp = 0000, $frac \neq 000$ the choices of $frac$ is from 001 to 111, in total $|frac|=7$.

Therefore, in total $1\times1\times7=7$.

i. (3 points) How many of these are negative, denormalized value?

Sol: Similar as above except s = 1.

Therefore, in total 7.

j. (3 points) What is the smallest positive normalized value?

Sol: $V=(-1)^s\times M\times 2^E$ where $s = 0$, $exp \neq 0000$ or 1111

for smallest positive normalized value, exp = 0001, $frac$ =000

Therefore, $M = 1+frac\times2^{-n} =1$; $E=exp-bias =1-7=-6$

$V=(-1)^s\times M\times 2^E =1\times1\times2^{-6}=\frac{1}{64}=\frac{8}{512}$

k. (3 point) What is the largest normalized value?

Sol: $V=(-1)^s\times M\times 2^E$ where $s = 0$, $exp \neq 0000$ or 1111

for largest positive normalized value, exp = 1110, $frac$ = 111

Therefore, $M = 1+frac\times2^{-n} = 1+ 111\times2^{-3}=1.111$; $E=exp-bias =14-7= 7$

$V=(-1)^s\times M\times 2^E =1\times1.111\times2^7= 11110000=240$

l. (3 points) What is the largest denormalized value?

Sol: $V=(-1)^s\times M\times 2^E$ where $s = 0$, exp = 0000

for largest denormailzed value, $frac$ = 111

Therefore, $M = frac\times2^{-n}= 111\times2^{-3}=.111$; $E=1-bias =1-7= -6$

$V=(-1)^s\times M\times 2^E =1\times.111\times2^{-6}= 111\times2^{-9}=\frac{7}{512}$

m. (3 points) What is the approximate number of decimal places of accuracy(i.e. significant figure)?

Sol: $n = 3$; 2^{-3} is about in the order of 10^{-1}. So the approximate number of decimal place of accuracy is 1.

2. (20 points) An integer 3,510,593 has hexadecimal representation 0x00359141, while the single-precision, floating-point number 3510593.0 has hexadecimal representation 0x4A564504. Derive this floating-point representation.

Sol: V = 3,510,593 = 0x 00359141

$V = 0000,0000,0011,0101,1001,0001,0100,0001$

$= 1.10101100100010100000 \times 2^{21}$

$= (-1)^s \times M \times 2^E$

$s = 0$; $E = 21 = exp - Bias \Rightarrow \text{exp} = E + Bias = 21+127=148 = 128+16+4$

$exp = 1001,0100$

$M = 1 + frac \times 2^{-23} = 1.f_{22}f_{21}\ldots f_1 f_0 = 1.10101100100010100000$

Hence $frac = 101,0110,0100,0101,0000,0100$

Its IEEE single-precision representation is

s	*exp*	*frac*
0	1001,0100	101,0110,0100,0101,0000,0100

which is 0, 1001,0100, 101,0110,0100,0101,0000,0100 which is regrouped in 4 bits

i.e., 0100,1010,0101,0110,0100,0101,0000,0100 in hexadecimal format

is 0x4A564504

3. (40 points) Given a floating-point format with a k-bit exponent and an n-bit fraction, write formulas for the exponent *E*, significand *M*, the fraction *f*, and the value *V* for the quantities that follow. In addition, describe the bit representation.

A. (10 points)The number 7.0

Sol: $V = 7.0 = 111 = 1.11 \times 2^2 = M \times 2^E$

$\Rightarrow M = 1.11$ and $E = 2$

$\Rightarrow M = 1 + .11 = 1 + frac \times 2^{-n}$

$\Rightarrow$ $frac = 1100\ldots0$ with n-2 bits of zeros.

$exp = E + Bias = 2 + 2^{k-1} - 1 = 2^{k-1} + 1 = 100\ldots01$ with k-2 bits of zeros

Its bit representation is

100…01, 1100…0

B. (15 points) The largest odd integer that can be represented exactly.

Sol: Assume *k*, *n* are fixed values.

$V = M \times 2^E = [1 + frac \times 2^{-n}] \times 2^E = [2^n + frac] \times 2^{(E-n)} = [2^n + frac] \times 2^{(exp-Bias-n)}$

$= [2^n + frac] \times 2^{exp-(2^{k-1}-1)-n}$

Let $T = exp - (2^{k-1} - 1) - n$, then $V = [2^n + frac] \times 2^T$

if $T \geq 1$, then 2^T is even, so is V, no matter what value *frac* is.

Therefore in order to have the largest odd value, T = 0 i.e., $exp - (2^{k-1}-1) - n = 0 \;\Rightarrow\; exp = 2^{k-1} + n-1$

exp = 1 xx…x with the significant bit 1 and its total value is $2^{k-1} + n-1$.

Also we want 2^n+frac to be the largest odd value, which means *frac* to be odd and the largest value for *frac* is 11…1 with *n* bits of 1.

Its bit representation is

exp	*frac*
1 xx…x	11…1
Its last *k*-1 bits value is *n*-1	All n bits be 1

C. (15 points) The reciprocal of the smallest positive normalized value.

Sol: smallest normalized value V_{min}

=> $exp = 00\ldots01 \Rightarrow E = 1 - Bias = 1 - (2^{k-1}-1) = 2-2^{k-1}$

and $frac = 00\ldots0 \Rightarrow M = 1 + frac\times 2^{-n} = 1$

$V_{min} = M\times 2^E = 1\times 2^E$

Let $V' = \frac{1}{V_{min}} = 2^{-E} = 2^{(2^{k-1}-2)} = M' \times 2^{E'}$

So $M' = 1.0 = 1 + frac' \times 2^{-n} \Rightarrow frac' = 00\ldots0$

$E' = exp' - Bias' \Rightarrow exp' = E' + Bias' = 2^{k-1}-2 + 2^{k-1}-1 = (2^k-1)-2$

Therefore $exp' = 11\ldots101$

Its bit representation is

11…101, 00…0