Maths 1263 Quiz 2 Friday 17/10/14

Your answers should show all work.

(1) Find the single-precision floating-point representation of $-88/9$. Give your answer in hexadecimal, little endian.

```
Answer: given - 88/9

sign 1 exponent 3  0111 1111
                          11
                   1000 0010
mantissa 11/9
11/9 4/9 8/9 16/9 14/9 10/9  2/9 4/9
   1 0   0   1    1    1     0   0

1.(001110)*

Verify
1 + 14 x (1/64) x (64/63) = 1 + 14/63 = 1+2/9.

mantissa
001110 001110 001110 00111 | 0...
round down.

Combining
1 1000 0010 001110 001110 001110 00111
1 100,0 001,0 001,110 0,0111,0 001,110 0,0111
c     1     1     c     7    1     c     7
little endian
c7 71 1c c1
```

In the remainder, we have invented a 10-bit floating-point number system with 1 sign bit, 4 biased exponent bits, and 5 mantissa bits.

(2) (a) What are N_{min} and N_{max} in this system? (b) What is $+\infty$? (c) What is -0? (d) What is ϵ_{mach}? (Give the answers to (a) and (d) as fractions; (b) and (c) can be binary or hex. ‘Little endian’ is not required.)

Answers.

$$
\begin{gathered}
\text{(a)} N_{\text{min}} = \frac{1}{64} \\
N_{\text{max}} = 252 \quad (2^7 \times 63/32) \\
\text{(b)} 0111100000 \\
\text{(c)} 1000000000 \\
\text{(d)} \frac{1}{32}
\end{gathered}
$$

In general, suppose there are e biased exponent bits and m mantissa bits.

- In the biased exponent, the minimum at face value is 0, and it should be used in representing zero. The maximum is $2^e - 1$, and it should be used for $\pm\infty$. Otherwise exponents are from 1 to $2^e - 2$ at face value. In biasing an exponent, $2^{e-1} - 1$ is added, so in reversing the effect, $2^{e-1} - 1$ is subtracted. Therefore the true exponent range is

$$
\begin{gathered}
-2^{e-1} + 1 \quad \text{for} \quad 0 \\
-2^{e-1} + 2 \ldots 2^{e-1} - 1 \quad \text{normal range} \\
2^{e-1} \quad \text{for} \quad \infty.
\end{gathered}
$$

- In the normal range, the minimum (true) mantissa possible is 1.0 (binary) and the maximum is $1.1 \ldots 1$ with m after the binary point. This is $2 - 2^{-m}$.

  The smallest nonzero value storable as an m-bit mantissa is $0 \ldots 01$, which represents $1.0 \ldots 1$ or $1 + 2^{-m}$ for normalised numbers.

  ϵ_{mach} is defined as: $1 + \epsilon_{\text{mach}}$ is the smallest normalised number greater than 1, so $\epsilon_{\text{mach}} = 2^{-m}$.

-

$$
\begin{gathered}
N_{\text{min}} = 2^{-2^{e-1}+2} \\
N_{\text{max}} = 2^{2^{e-1}-1}(2 - 2^{-m}) \\
\epsilon_{\text{mach}} = 2^{-m}
\end{gathered}
$$

Thus, with $e = 4$ and $m = 5$,

$$
\begin{gathered}
\infty = 2^8 \\
N_{\text{min}} = 2^{-8+2} = 1/64 \\
N_{\text{max}} = 2^7(2 - 2^{-5}) = 252 \\
\epsilon_{\text{mach}} = 1/32.
\end{gathered}
$$

(3) Convert $33/16$ and $39/64$, which are floating-point numbers in this system, into the form

$$1.b_1b_2b_3b_4b_5 \ \times \ 2^e.$$

Also calculate (exactly) their sum and difference as proper fractions.
Be careful: the correct answers are needed below.

```
Answers.
33/16 = 33/32 x 2^1 = 1.00001 x 2^1
39/64 = 39/32 x 2^{-1} = 1 7/32 x 2^{-1} = 1.00111 x 2^{-1}
sum: (132+39)/64 = 171/64
difference: (132-39)/64 = 93/64
```

(4) Add the two 10-bit floating-point numbers in (3), correctly rounded, using 9 bits (6-bit significand, possibly shifted, and the G,R,S bits).

```
Answer.
1.00001 x 2^1
1.10111 x 2^{-1}   Shift....

 1.00001 00 x 2^1
+0.01001 11 x 2^1 G 1 R 1 S 0
 -------------------------
 1.01010 11 round up ....

 1.01011 x 2^1 --- answer
```

(this is $43/16 = 172/64$).

(5) Subtract the smaller 10-bit floating-point number in (4) from the larger, correctly rounded, using 9 bits.

```
Answer.
 1.00001 00 x 2^1
-0.01001 11 x 2^1 G 1 R 1 S 0
-------------------
 0.10111 01 left shift
 1.01110 1 and S=0; round to evens; exponent is now 0
 1.01110 x 2^0 --- answer
```

(this is $46/32 = 92/64$).