Homework 1

Integer and Floating Point Number Representations

Integer

Problem 1

Suppose you have a 2 GHz x64 core and you can execute three integer operations (additions or subtractions) every cycle. How long (how many seconds) will the following loop run?

```
uint32_t i;   /* 32 bit unsigned integer */
uint64_t s;   /* 64 bit unsigned integer */
for (i = 1 ; i != 0; i++) {
      s += i;
}
```

Problem 2

How can you compute the following using only shifts, adds, and subtracts? Here, `x` is a `uint32_t`.

```
16 * x
17 * x
23 * x
x / 16
x / 17  (Hard! Outside the scope of class, but included as a challenge)
```

Problem 3

Some instruction sets, including x64, provide an integer representation in addition to two's complement. This representation is called Binary Coded Decimal (BCD). In BCD, a decimal digit (0,1,2,3,4,5,6,7,8,9) is encoded into a group of 4 bits using 0000 through 1001. How many unique numbers can be represented in a 64 bit BCD quantity? Why might one use BCD to represent prices like $10.99 instead of using fixed point binary or floating point?

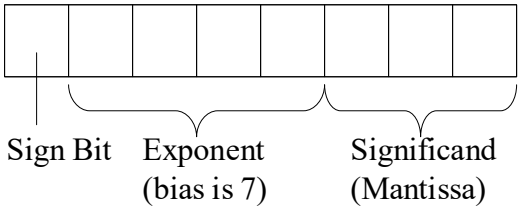
Problem 4

Many instruction sets have instructions where when you multiply two k bit numbers the result is stored as two k bit numbers. Why? Similarly, many have instructions where if you divide two k bit numbers, the result is stored as two k bit numbers. Why? By "k bit numbers" think of unsigned integers, although the same reasoning holds for signed integers.

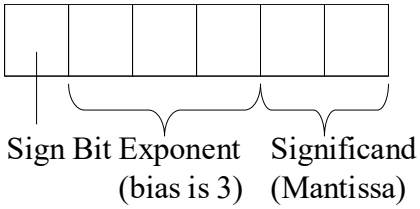
Floating Point

Consider the following two small floating point formats based on the IEEE standard:

- Little Format



- Tiny Format



Except for the sizes of these formats, the rules are those of the IEEE standard.

Problem 1

For both formats, determine the following values (in decimal)

1. Largest positive finite number
2. Positive normalized number closest to zero
3. Largest negative denormalized number
4. Negative denormalized number closest to zero

Problem 2

Encode the following values in the 8 bit Little Format: ¾, -13/16, 44, –104, NaN, and negative infinity. Show each in binary and hexadecimal.

Problem 3

Determine the values corresponding to the following Little Format bit patterns. The leftmost bit is the most significant

1. 10101011
2. 01111000
3. 10110101
4. 01011111
5. 11000101
6. 11111111

Problem 4

Convert the following 8 bit Little Format numbers into 6 bit Tiny Format numbers. Overflow should yield +/- infinity, underflow should yield +/- 0.0, and rounding should follow the “round-to-nearest-even” tie-breaking rule. Show the bit pattern and its hex representation.

1. 00010010
2. 11101011
3. 10100011
4. 11001110
5. 00110101
6. 11111111
7. 01111000

Problem 5

The changing demands of scientific computation, the growing importance of machine learning computation, and issues with the IEEE floating point standard you are learning have resulted in a range of new alternatives that current vying for attention.[1] For machine learning purposes (specifically neural network "deep learning" computations), it is generally believed that 16 bit floating point is sufficient.

- A 16 bit version of the IEEE standard exists and is widely implemented. A float16 has 5 exponent bits and 10 mantissa bits.
- Brain Floating Point is another important standard (also widely implemented in recent Intel and ARM chips, and originating in Google's TPU chips). A bfloat16 has 8 exponent bits and 7 mantissa bits. It is essentially a 32 bit IEEE floating point number with 16 bits chopped off the end of the mantissa.

What are the comparative advantages and disadvantages of these formats in terms of the real numbers they can represent?

Some proposed alternatives to IEEE floating point encode the split between exponent bits and mantissa bits directly in the number itself. For example, for an n bit number, we might introduce a log2(n) bit field that encodes the position of the last exponent bit. For 16 bits, this field would be 4 bits wide, leaving just 12 bits for the sign, exponent and mantissa. For 32 bits, the field would be 5 bits wide, and so on. What are the advantages and disadvantages of such a scheme? What happens as the bit width of the number increases?

Problem 6

The floating point (and integer, fixed point, bcd, etc) numbers and their operators (+/-/*/…) are generally implemented in hardware for speed. However, software implementations also exist, such as GNU MPFR. While these are much slower, they allow a programmer to use a bit width that the hardware does not support by emulating it using software. For example, the programmer could choose to use 1024 bit floating point numbers or 2048 bit integers. Why might a programmer do this?

(Thought experiment – nothing to hand in) Could you write a program that implements a rational number system? In a rational number system, every value is represented as a fraction (a ratio of two integers).

[1] Other important examples: unums and posits