Fall 2017 Math 395 Written Homework 2 Key 100 total. -5 for no stapling

1. Convert binary to decimal.

   (a) 110_2

   (b) $11.0101_2 = 3 + 0.25 + 0.0625 = 3.3125$

   Solution: (a) $2^2 + 2^1 = 6$

   (b) $2 + 1 + 2^{-2} + 2^{-4}$

2. Directly add and subtract in binary (No rounding involved).

   (a) $0.111_2 + 11.101_2$

   (b) $1.0101_2 - 1.10_2$

   (c) $1.1010_2 \times 2^{-4} + 1.1010_2 \times 2^{-3}$. Convert it to decimal with a calculator.

   Solution: (a) 100.100_2

   (b) -0.0011_2

   (c) $1.1010_2 \times 2^{-4} + 1.1010_2 \times 2^{-3} = 0.11010_2 \times 2^{-3} + 1.1010_2 \times 2^{-3} = 10.0111_2 \times 2^{-3} = 1.00111_2 \times 2^{-2} = (1 + 2^{-3} + 2^{-4} + 2^{-5})2^{-2} = 0.3046875$

3. Compute the **normalized** floating point numbers

   (a) 10.3456, base 10, retain 4 digits after point, rounding up.

   (b) 1.0101_2, base 2, retain 2 digits after point, rounding up.

   (c) 1.0101_2, base 2, retain 2 digits after point, rounding down.

   (d) 1.0101_2, base 2, retain 2 digits after point, rounding to the nearest (decide which one is closer by finding the midpoint of answers in (b) and (c)).

   Solution: (a) 1.0346×10

   (b) 1.10_2

   (c) 1.01_2

   (d) Midpoint of 1.10_2 and 1.01_2 is 1.011_2. 1.0101_2 is smaller than the midpoint, so rounding to the nearest is rounding down here: 1.01_2

4. Compute the machine result of 0.1+0.2 if we retain 2 digits precision after the binary point.

   Solution: $0.1 = 1.10011001100\cdots_2 \times 2^{-4}$. Rounding down $D(0.1) = 1.10_2 \times 2^{-4}$. Rounding up $U(0.1) = 1.11_2 \times 2^{-4}$. Midpoint $\dfrac{D(0.1) + U(0.1)}{2} = 1.101_2 \times 2^{-4} > 1.1001100\cdots_2 \times 2^{-4}$

   $fl(0.1) = D(0.1) = 1.10_2 \times 2^{-4}$. Similarly, $fl(0.2) = 1.10_2 \times 2^{-3}$.

   Now we make both numbers have power 2^{-3}, then add, then normalize, then round:

   $fl(0.1) \oplus f(0.2) = fl(0.11_2 \times 2^{-3} + 1.10_2 \times 2^{-3}) = fl(10.01_2 \times 2^{-3}) = fl(1.001_2 \times 2^{-2}) = 1.00_2 \times 2^{-2} = 0.25$.

5. *In Python

```
>>>  (0.1 + 0.2) + 0.3
0.6000000000000001
>>>  0.1 + (0.2 + 0.3)
0.6
```

Explain this phenomenon with a more limited machine where we only retain 2 digits after the binary point.

Solution: $0.3 = 2^{-2} + 2^{-5} + 2^{-6} + \cdots = 1.0011\cdots_2 \times 2^{-2}$.

Rounding down: $D(0.3) = 1.00_2 \times 2^{-2}$

Rounding up: $U(0.3) = 1.01_2 \times 2^{-2}$

midpoint $1.001_2 \times 2^{-2} < 0.3$, so rounding up is nearer: $fl(0.3) = 1.01_2 \times 2^{-2}$.

By Exercise 4, $0.1 \oplus 0.2 = 1.00_2 \times 2^{-2}$

$(0.1 \oplus 0.2) \oplus 0.3 = fl(1.00_2 \times 2^{-2} + 1.01_2 \times 2^{-2}) = fl(10.01 \times 2^{-2}) = fl(1.001_2 \times 2^{-1}) = 1.00_2 \times 2^{-1} = \boxed{0.5}$

$0.2 \oplus 0.3 = fl(0.11_2 \times 2^{-2} + 1.01_2 \times 2^{-2}) = fl(1.000_2 \times 2^{-1}) = 1.00_2 \times 2^{-1}$.

$0.1 \oplus (0.2 \oplus 0.3) = fl(1.10_2 \times 2^{-4} + 1.00_2 \times 2^{-1}) = fl(0.00110_2 \times 2^{-1} + 1.00_2 \times 2^{-1}) = fl(0.01_2 \times 2^{-1} + 1.00_2 \times 2^{-1}) = fl(1.01_2 \times 2^{-1}) = 1.01_2 \times 2^{-1} = \boxed{0.625}$

6. In the 1991 Gulf War, the Patriot misslie defense system failed due to roundoff error. The troubles stemmed from a computer that performed the tracking calculations with an internal clock whose integer values in tenths of a second were converted to seconds by multiplying by a 24-bit binary approximation to 0.1:

$$0.1 \approx 0.000110011001100110011001_2$$

   (a) Convert the binary number above to a fraction. Call it x.

   (b) Compute the absolute difference between 0.1 and x.

   (c) What is the time error in seconds after 100 hours of operation (i.e., the value of $|360,000 - 3,600,000x|$)?

On February 25, 1991, a Patriot battery system, which was to protet the Dhahran Air Base, had been operating for over 100 consecutive hours. The roundoff error caused the system not to track an incoming Scud missle, which slipped through the defense system and detonated on US Army barracks, killing 28 American soldiers.

Solution: (a) $x = \dfrac{2^{17} + 2^{16} + 2^{13} + 2^{12} + 2^9 + 2^8 + 2^5 + 2^4 + 2 + 1}{2^{21}} = \dfrac{209715}{2097152}$

(b) $|0.1 - x| = \dfrac{1}{10} - \dfrac{209715}{2097152} = \dfrac{2}{20971520} \approx 9.53674 \times 10^{-8}$.

(c) $|360,000 - 3,600,000x| = 3,600,000|0.1 - x| \approx 0.34$ seconds