Numerical Analysis MATH50003 (2023–24) Problem Sheet 3

Problem 1 With 8-bit unsigned integers, what is the result for the following computations:

$$127 \oplus_{256} 200, \quad 2 \otimes_{256} 127, \quad 2 \otimes_{256} 128, \quad 0 \ominus_{256} 1$$

SOLUTION

$$127 \oplus_{256} 200 = 327 \ (\text{mod} 256) = 71$$

$$2 \otimes_{256} 127 = 254$$

$$2 \otimes_{256} 128 = 256 \ (\text{mod} 256) = 0$$

$$0 \ominus_{256} 1 = -1 \ (\text{mod} 256) = 255.$$

END

Problem 2(a) With 8-bit signed integers, what are the bits for the following: $10, 120, -10$.

SOLUTION We can find the binary digits by repeatedly subtracting the largest power of 2 less than a number until we reach 0, e.g. $10 - 2^3 - 2 = 0$ implies $10 = (1010)_2$. Thus the bits are: `00001010`. Similarly,

$$120 = 2^6 + 2^5 + 2^4 + 2^3 = (1111000)_2$$

Thus the bits are `01111000`. For negative numbers we perform the same trick but adding 2^p to make it positive, e.g.,

$$-10 = 2^8 - 10(\text{mod} 2^8) = 246 = 2^7 + 2^6 + 2^5 + 2^4 + 2^2 + 2 = (11110110)_2$$

This the bits are: `11110110`. **END**

Problem 2(b) With 8-bit signed integers, what is the result for the following computations:

$$127 \oplus_{256}^{\rm s} 200, \quad 2 \otimes_{256}^{\rm s} 127, \quad 2 \otimes_{256}^{\rm s} 128, \quad 0 \ominus_{256}^{\rm s} 1$$

SOLUTION

$$127 \oplus_{256}^{\rm s} 200 = 327 \ (\text{mod}^{\rm s} 256) = 71$$

$$2 \otimes_{256}^{\rm s} 127 = 254 \ (\text{mod}^{\rm s} 256) = -2$$

(The third part was a trick question: 128 cannot be represented as an 8-bit signed integer)

$$0 \ominus_{256}^{\rm s} 1 = -1 \ (\text{mod} 256)^{\rm s} = -1.$$

END

Problem 3 What is π to 5 binary places? Hint: recall that $\pi \approx 3.14$.

SOLUTION We subtract off powers of two until we get 5 places. Eg we have

$$\pi = 3.14\ldots = 2{+}1.14\ldots = 2{+}1{+}0.14\ldots = 2{+}1{+}1/8{+}0.016\ldots = 2{+}1{+}1/8{+}1/64{+}0.000\ldots$$

Thus we have $\pi = (11.001001\ldots)_2$. The question is slightly ambiguous whether we want to round to 5 digits so either `11.00100` or `11.00101` would be acceptable. **END**

Problem 4 What are the single precision $F_{32} = F_{127,8,23}$ floating point representations for the following:

$$2, \quad 31, \quad 32, \quad 23/4, \quad (23/4) \times 2^{100}$$

SOLUTION Recall that we have σ,`Q,S = 127,8,23`. Thus we write

$$2 = 2^{128-127} * (1.00000000000000000000000)_2$$

The exponent bits are those of

$$128 = 2^7 = (10000000)_2$$

Hence we get the bits

```
0 10000000 00000000000000000000000
```

We write

$$31 = (11111)_2 = 2^{131-127} * (1.1111)_2$$

And note that $131 = (10000011)_2$ Hence we have the bits

```
0 10000011 11110000000000000000000
```

On the other hand,

$$32 = (100000)_2 = 2^{132-127}$$

and $132 = (10000100)_2$ hence we have the bits

```
0 10000100 00000000000000000000000
```

Note that

$$23/4 = 2^{-2} * (10111)_2 = 2^{129-127} * (1.0111)_2$$

and $129 = (10000001)_2$ hence we get:

```
0 10000001 01110000000000000000000
```

Finally,

$$23/4 * 2^{100} = 2^{229-127} * (1.0111)_2$$

and $229 = (11100101)_2$ giving us:

```
0 11100101 01110000000000000000000
```

END

Problem 5 Let $m(y) = \min\{x \in F_{32} : x > y\}$ be the smallest single precision number greater than y. What is $m(2) - 2$ and $m(1024) - 1024$?

SOLUTION The next float after 2 is $2 * (1 + 2^{-23})$ hence we get $m(2) - 2 = 2^{-22}$:

```
nextfloat(2f0) - 2, 2^(-22)
```

```
(2.3841858f-7, 2.384185791015625e-7)
```

similarly, for $1024 = 2^{10}$ we find that the difference $m(1024) - 1024$ is $2^{10-23} = 2^{-13}$:

```
nextfloat(1024f0) - 1024, 2^(-13)
```

```
(0.00012207031f0, 0.0001220703125)
```

END