CSCI 247

Isaac Boaz

January 18, 2023

Question 1

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

A) <u>The number 7.0</u>

Keeping in mind the general formula

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

where V is the value,
s is the sign,
M is the mantissa,
and E is the exponent.

1. We'll first begin by converting 7.0 to its binary representation
   $7_{10} = 111_2$
2. Normalizing this representation leaves us with

$$111_2 = 1.11 \cdot 2^2$$

3. Meaning the exponent $Exp = 2$
4. And the fractional part is $f = 111^*$
   Note: IEEE implies a leading 1 at the beginning, so $f = 110$
5. Calculate the biasd E by adding $Bias = 2^{k-1} - 1$

$$E = 2 + (2^{k-1} - 1) = 2^{k-1} + 1$$

6. Combing with an example of $k = 4, M = 3$

$$\begin{aligned} Exp &= 2 \\ E &= 2^{4-1} + 1 = 9_{10} = 1001_2 \\ M &= 1.11_2 \\ f &= 110_2 \\ V &= (-1)^s \cdot (1.)f \cdot 2^{Exp} \\ &= (1.)11_2 \cdot 2^2 = 7.0 \end{aligned}$$

With a final bit representation of

Sign	Exp	Mantissa
0	1001	110

B) The largest odd integer that can be represented exactly

- We know the sign must be 0 for positives.
- The mantissa should be all 1's to get the largest representable odd number.
- If our exponent goes beyond the mantissa's accuracy, we'll end up with a 0 as the least significant bit (IE an even number).
- Thus, our exponent is $Exp = Min(k, M)$

An example where $k = 4, M = 3$

$$\begin{aligned}
Exp &= 3 \\
Bias &= 2^{k-1} - 1 = 2^{4-1} - 1 = 7 \\
E &= Exp + Bias \\
&= 3 + 7 = 10_{10} = 1010_2 \\
f &= 111_2 \\
V &= (1.)111_2 \cdot 2^3 = 15
\end{aligned}$$

With a final bit representation of

Sign	Exp	Mantissa
0	1010	111

C) The reciprocal of the smallest positive normalized value

In order to find the reciprocal of any number, we can simply "flip" the numerator and denominator. This can be achieved in floating point representation by multiplying the exponent by -1.
The smallest possible value given k exponent bits and M fraction bits is when:

- ☐ $M = 0,\ s = 0$
- ☐ Exp is the biggest positive value multiplied by -1

1. Set $M = 0,\ s = 0$
2. Calculate $E = Exp - Bias$ with $Exp = 1$
3. One example where $k = 5, M = 4$

$$\begin{aligned}
Bias &= 2^{k-1} - 1 = && 2^{5-1} - 1 && = 15 \\
E &= (1 - Bias) = && (1 - 15) \cdot -1 && = 14 \\
Exp &= E + Bias = && (14 + 15) && = 29_{10} = 11101_2 \\
f &= && && 0_2
\end{aligned}$$

With a final bit representation of

Sign	Exp	Mantissa
0	11101	000

Which equals 16384