CSCI 2330 – Floating Point Exercises

1. Using our toy 8-bit floating point representation (with k=4 exponent bits and 3 fractional bits), convert **00110100** into a decimal value.

2. Using the same 8-bit representation, convert **10000101** into a decimal value (working with a fraction here is advisable).

3. What is the decimal value of the largest possible 32-bit IEEE floating point number (excluding infinity)? You do not need to simplify the expression but should be able to write down the exact value.

4. If **d** is a double in C, does d < 0.0 imply ((d * 2) < 0.0)?

5. You may wonder why IEEE 754 encodes the exponent value **E** using an unsigned **exp** field and a **bias** value instead of just making **exp** a signed int. Why might the designers have made this choice?

(*Hint*: one of the design goals of IEEE 754 was to have floating point numbers ordered in the same way as if they were ints, to allow for easy comparisons -- e.g., 001 < 010 < 011 regardless of whether they are ints or floating point. Don't worry about negative values here.)