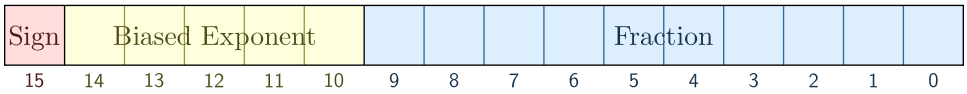
Homework 7

The IEEE 754 standard was updated in 2008 to include a half-width floating-point format in addition to the standard-width, double-width, and other formats. It looks like this:



- The entire number is only 16 bits wide.
- The exponent field is 5 bits wide, so the bias is -15.
- The fraction field is 10 bits wide.

Clearly, this "half float" is more efficient than the other standards, but the numbers are not as accurate. It is most commonly seen in graphics hardware, such as that made by Nvidia. Problems 1-3 make use of this data type.

(Hint: if your calculator or phone cannot handle fractional binary arithmetic, *please find an online calculator that does*. It will make this assignment much more approachable.)

1. **A.** Use half floats to calculate both $(-16{,}360 + 16{,}360) + 1$ and $-16{,}360 + (16{,}360 + 1)$. Does associativity hold in this case? *(1.0 (0 01111 0000000000); 0.0 (0 00000 0000000000); no)*

B. Use half floats to calculate both $(28.65625 + 0.4140625) + 12.140625$ and $28.65625 + (0.4140625 + 12.140625)$. Does associativity hold in this case? *(41.1875 (0 10100 0100100110); 41.21875 0 10100 0100100111; no)*

2. **A.** Use half floats to calculate both $(0.00048828125 \times 1768) \times 250.125$ and $0.00048828125 \times (1768 \times 250.125)$. Does associativity hold in this case? *(215.875 (0 10110 1010111111); ∞ (0 11111 0000000000); no)*

B. Use half floats to calculate both $(47.21875 \times 28.09375) \times 35.75$ and $47.21875 \times (28.09375 \times 35.75)$ Does associativity hold in this case? *(47456 (0 11110 0111001011); 47424 (0 11110 0111001010); no)*

3. **A.** Use half floats to calculate both $0.15234375 \times (0.20703125 + 99.6875)$ and $(0.15234375 \times 0.20703125) + (0.15234375 \times 99.6875)$. Does the distributive property hold in this case? *(both 15.21875 (0 10010 1110011100); yes)*

B. Use half floats to calculate both $-27.890625 \times (-8088 + 10216)$ and $(-27.890625 \times -8088) + (-27.890625 \times 10216)$. Does the distributive property hold in this case? *(-59360 (1 11110 1100111111); NaN (0 11111 ??????????); no)*

4. **A.** Using the standard 32-bit floating-point number, represent $1/3$ as accurately as possible. Is your answer exact? *(0.33333334326744079589843 75, or 0 01111101 01010101010101010101011; no)*

B. What is $(1/3 + 1/3 + 1/3)$ using these standard floats? Be as exact as possible. *(1.0, or 0 01111111 00000000000000000000)*

C. What is $(1 - 1/3 - 1/3 - 1/3)$ using these standard floats? Be as exact as possible. *(-5.960464477539062 5 $\times 10^{-8}$, or 1 01100111 00000000000000000000)*