cs281: Introduction to Computer Systems

Floating Point Table Example

Assume we are using the IEEE floating point format with 5 bits used as: 1 bit for the sign, 2 bits for the exponent ($k = 2$) and 2 bits for the fraction ($n = 2$).

1. First draw a picture of the bit vector for this scheme. Label each part.
2. Compute the bias term.
3. Now complete the table on the back of this page.
4. Draw horizontal lines separating the three cases: Normalized, Denormalized, and Special.
5. What is the bit representation for zero?
6. What is the smallest value possible which is greater than zero?
7. What is the largest value possible with value less than 1?
8. What is the bit representation for the value of 1?
9. What is the largest possible value (non special case)?
10. What is the bit representation for $+\infty$?
11. Draw a number line similar to Figure 2.33 (page 106) in your text and plot the points from your table on the number line.

sign	exp		frac		e	E	2^E	f	M	M(bits)	$M * 2^E$ (bits)	value
0	0	0	0	0								
0	0	0	0	1								
0	0	0	1	0								
0	0	0	1	1								
0	0	1	0	0								
0	0	1	0	1	1	0	1	$\frac{1}{4}$	$\frac{5}{4}$	1.01	1.01	1.25
0	0	1	1	0								
0	0	1	1	1								
0	1	0	0	0								
0	1	0	0	1								
0	1	0	1	0								
0	1	0	1	1								
0	1	1	0	0								
0	1	1	0	1								
0	1	1	1	0								
0	1	1	1	1								