Example

The question I was trying to do on the overhead was:

> **With 5 bits for s, 4 bits for e, two guard bits and a sticky bit: write $3/16$ and $1/3$ in floating point format and calculate their sum.**

The numbers will be in the form:

$$\pm 2^{e-7}(1.s)$$

Where does the 7 come from? Well there are 4 bits for e, so e ranges from 0 to 15. To center the range of possible exponents we subtract half of 15, which is 7 when rounded down.

First we write 3/16 in this form:

$$3/16 = 3 \times 2^{-4} = 1.5 \times 2^{-3} = 2^{4-7}(1.10000)_2$$

So $e = 0100$ and $s = 10000$. The bit pattern for 3/16 will be 0010010000.

By long division $1/3 = (0.01010101\ldots)_2$.

$$1.010101\ldots 2^{-2} = 2^{5-7}(1.01011)_2$$

Guard bits are 10, so we don't know if we should round up or down. The sticky bit is 1 as there are non zero bits to right, so round up. $e = 0101, s = 01011$. The bit pattern for 1/3 will be 0010101011.

Now we align the numbers. The second number (1/3) has a larger magnitude so we shift the first number left by the difference in their exponents. Since the sign is the same we can just add.

$$\begin{array}{l} 2^{5-7}(00.110000)_2 \\ 2^{5-7}(01.01011)_2 \\ \hline 2^{5-7}(10.000110)_2 \end{array}$$

Normalising this we get $2^{6-7}(1.0000110)_2$. The guard bits are 10 and the sticky bit is zero, so we truncate to $2^{6-7}(1.00001)_2$. So $e = 0110$ and $s = 00001$. Final bit pattern 0011000001.