

Here is an example of why you need a Guard bit, in addition to the Round and Sticky bits.

Consider the subtraction problem below, first with **infinite precision**, then with Guard, Round and Sticky bits (like in IEEE FP), and finally just using Round and Sticky bits (**not correct**).

$$\begin{array}{r} 1.000 \times 2^5 \\ - 1.001 \times 2^1 \\ \hline \end{array}$$

With infinite precision, we **keep ALL the bits** as we shift the fraction right to make the exponents equal. Note that we ALWAYS shift right, to make the logic easier to implement. After we perform the subtraction, we need to shift the result left one bit to renormalize the result (because the result has a leading 0). Then we need to round the result to have (in this example) three bits to the right of the radix point. We call the least significant bit of the rounded result the “unit of the last place” or ULP. If the bits to the right of the ULP have weight $>1/2$ ULP, then we need to round up. If they bits have weight $<1/2$ ULP, then we round down. And if the bits are exactly equal to $1/2$ ULP, then we round to even. In this example, there are three bits to the right of the ULP. If these three bits are 0xx, then we round down; if the bits are 100 we round to even; if the bits are 1xx and at least one of the x’s is a one, we round up.

$$\begin{array}{r} 1.000\ 0000 \times 2^5 \\ - 0.000\ 1001 \times 2^5 \\ \hline 0.111\ 0111 \times 2^5 \text{ Need to shift left to } \text{normalize} \\ 1.110\ 111 \times 2^4 \text{ Round up, since } \text{more than half} \text{ unit of the last place} \\ 1.111 \times 2^4 \end{array}$$

We can use the Guard, Round, and Sticky bits to reduce round the result correctly using on 3 extra bits in our adder/subtractor. As we shift the operand right, we shift the bits into the Guard, then the Round, and finally the Sticky bits. The Guard and Round bits are just standard bits, but the Sticky bit is 1 if ANY bit that **shifts through it is a 1**. In this example, the Sticky bit is set to 1 since **the first bit that shifts into it is a 1**. Note that the initial result of the computation is slightly different than the infinite precision version, but that **the rounded result is the same**.

$$\begin{array}{r} 1.000\ 000 \times 2^5 \\ - 0.000\ 101 \times 2^5 \text{ Guard is 1, Round 0, and Sticky is 1} \\ \hline 0.111\ 011 \times 2^5 \text{ Need to shift left to normalize, using Guard bit} \\ 1.110\ 11 \times 2^4 \text{ Round up, since more than half unit of the last place} \\ 1.111 \times 2^4 \text{ Result is correctly rounded} \end{array}$$

It isn't always obvious why you need the **Guard** bit, and can't get by with just the Round and Sticky bits. But this example shows you why you need the Guard bit. If we do the calculation with just the Round and Sticky bits, we **use up the Round bit** when we have to **normalize the result** (since the bit to the left of the radix point is 0). Thus we are left with only the Sticky bit **to do the rounding**. But we can't really use the Sticky bit to round, because we can't tell whether it represents 001, **100**, or 111, which would all round differently.

$$\begin{array}{r}
 1.000\ 00 \times 2^5 \\
 - 0.000\ 11 \times 2^5 \quad \text{Round is 1, Sticky is 1} \\
 \hline
 0.111\ 01 \times 2^5 \quad \text{Need to shift left to normalize, must use Round bit} \\
 1.110\ 1 \times 2^4 \quad \text{Can't round using Sticky, since can't tell if } \geq / < 1/2 \text{ ULP}
 \end{array}$$

If it still isn't clear WHY you can't use the Sticky bit to round, consider the following SLIGHTLY DIFFERENT computation:

$$\begin{array}{r}
 1.000 \times 2^5 \\
 - 1.111 \times 2^1
 \end{array}$$

With infinite precision, the calculation is:

$$\begin{array}{r}
 1.000\ 0000 \times 2^5 \\
 - 0.000\ 1111 \times 2^5 \\
 \hline
 0.111\ 0001 \times 2^5 \quad \text{Need to shift left to normalize} \\
 1.110\ 001 \times 2^4 \quad \text{Round down, since } < 1/2 \text{ ULP} \\
 1.110 \times 2^4
 \end{array}$$

With Guard, Round and Sticky, the calculation round correctly:

$$\begin{array}{r}
 1.000\ 000 \times 2^5 \\
 - 0.000\ 111 \times 2^5 \\
 \hline
 0.111\ 001 \times 2^5 \quad \text{Need to shift left to normalize} \\
 1.110\ 01 \times 2^4 \quad \text{Round down, since less than half unit of the last place} \\
 1.110 \times 2^4
 \end{array}$$

With only Round and Sticky, the calculation becomes:

$$\begin{array}{r}
 1.000\ 00 \times 2^5 \\
 - 0.000\ 11 \times 2^5 \\
 \hline
 0.111\ 01 \times 2^5 \quad \text{Need to shift left to normalize} \\
 1.110\ 1 \times 2^4 \quad \text{Can't round using Sticky, since can't tell if } \geq / < 1/2 \text{ ULP}
 \end{array}$$

But notice that this last calculation is EXACTLY the same as the one we did before using just Round and Sticky, but in that case the right answer is to

round UP, and in this case the right answer is to round DOWN. Thus, we need the Guard bit to round correctly.