Role of Guard Bits and Rounding in Precision of Floating-Point Arithmetics

BIDYUT KUMAR PATRA

February 2020

Few bits called *guard bits* are padded to left of the mantissa to achieve better precision while performing arithmetic operations such as addition, subtraction, multiplication and division. Rounding is another policy that takes significant role in achieving better precision. These two aspects are described with an example.

Let X and Y be two numbers, where $X = 1.000 \times 2^5$ and $Y = 1.001 \times 2^1$. We will perform $Z = X - Y$ in three situations (scenario) described next.

- **Scenario 1(Infinite Precision):** We perform $X - Y$ considering the fact that we have registers of infinite length (Table 1).
  $Z = X - Y = 1.110\ 1110 \times 2^4 = 29.75$

$X = 1.000$	0000×2^5
$Y = 0.000$	1001×2^5
$Z = 0.111$	0111×2^5
$Z = 1.110$	1110×2^4

Table 1: With registers of infinite length

- **Scenarios 2 (Guard Bits with Rounding):** We use four guard bits and result is normalized. Finally, number is rounded to the closest re-presentable number. One can perform rounding in three ways.

  1. **Rounding Up:** If the value of the guard bits is more than the half of last re-presentable bit position, then result is **rounded up** by adding 1 at the least significant position. Let 1010 be the guard bits. The value of these bits is more than the half of the last re-presentable bit position (least significant bit position). In this example, weight of least significant bit position is 2^{-3} as mantissa is 3 bits only. If the value of the guard bits is more than the last re-presentable bit position, then result is rounded up by adding 1 at the least significant

position, i.e, 0.001 is added to 1.110 (Table 2). The value of $Z(X-Y)$ is $1.111 \times 2^4 = 30_{10}$. The error is $|29.75 - 30.00| = 0.25$

2. **Rounding Down:** If the value of the guard bits is lesser than the half of last re-presentable bit position, then all guard bits are truncated. Let 0110 be the guard bits. The value of these bits is less than the half of the last re-presentable bit position (least significant bit position). Therefore, guard bits will be truncated.
3. **Rounding to Even Number:** If the value of the guard bits is exactly the half of last re-presentable bit position, then 1 is added at the least significant bit position if normalized result has 1 at the significant position, ignore otherwise.

$X = 1.000$	0000×2^5	
$Y = 0.000$	1001×2^5	
$Z = 0.111$	0111×2^5	
$Z = 1.110$	1110×2^4	
$+ \quad 0.001$		
$Z = 1.111$	$\times 2^4$	$= 30_{10}$

Table 2: Guard Bits with Rounding

- **Scenarios 3 (Guard Bits without Rounding policy):** Guard bits are used to hold the bits while adjusting the exponent and during normalization. However, the value of the guard bits are not utilized for rounding. In this running example, Z will have value $1.110^4 = 28$.. The error is $|29.75 - 28| = 1.75$.
- **Scenario 4 (Without Guard bits):** If we do not use any guard bits, then Z will be $1.000 \times 2^5 = 32$. (Table 3). The error is $|29.75 - 32| = 2.75$.

$X = 1.000$	$\times 2^5$	
$Y = 0.000$	$\times 2^5$	
$Z = 1.000$	$\times 2^5$	$= 32$

Table 3: Without Guard Bits

From previous example, we conclude that with guard bits and rounding, we can reach close to the exact result (infinite precision). Question needs to be asked that how many guard bits are sufficient to get minimum error. IEEE 754 format suggests three bits namely, Guard bit (G), Round bit (R) and Sticky bit (S). The G and R bits act as general guard bits where as the S bit holds the value 1 if any 1 passes over it (Table 4). In this example, the S bit gets the value 1 while adjusting the exponent of smaller number and it remains 1 afterwards. However, the value of guard bits is less than the half of the least significant bit position ($011 < 0.5 \times 2^{-3}$). Therefore, rounding is not required. The sticky bit

helps to identify whether value of the guards is more than the half of the last bit position.

Consider a scenario where 100001 are the bits we receive from the right end of a mantissa while adjusting exponent of a smaller number. If we use three bits as the guard bits without using one as sticky bit, then value of the guard bits (100) will be exactly equal to the half of the last re-presentable bit position. However, if least significant bit of the three guard bits acts as sticky bit, then the value of the guard will be more than half of the last re-presentable bit position. Therefore, sticky bit plays an important role in achieving better precision.

	GRS	
$X = 1.000$	000×2^5	
$Y = 0.000$	101×2^5	
$Z = 0.111$	011×2^5	
$Z = 1.111$	$\times 2^4$	
$Z = 1.111$	$\times 2^4$	$= 30_{10}$

Table 4: With GRS (Guard, Round and Sticky)