Slide-13-16

$$\begin{array}{r} A = a_{n-1}\ a_{n-2} \cdots a_0 \\ B = b_{n-1}\ b_{n-2} \cdots b_0 \\ \hline S = s_{n-1}\ s_{n-2} \cdots s_0 \end{array}$$

(c_n and c_{n-1} are the carries into and out of the sign bit)

Case I: $c_{n-1}=0,\ c_n=1$

$\Rightarrow a_{n-1}=b_{n-1}=1,\ s_{n-1}=0$

$\Rightarrow$ overflow because adding two negative numbers cannot result in a positive number.

Case II: $c_{n-1}=1,\ c_n=0$

$\Rightarrow a_{n-1}=b_{n-1}=0,\ s_{n-1}=1$

$\Rightarrow$ overflow because adding two positive numbers cannot result in a negative number

Case III: $c_{n-1}=c_n=0$

$\Rightarrow$ at most one of a_{n-1} and b_{n-1} is 1.

Case III A: $a_{n-1}=b_{n-1}=0$

$\Rightarrow$ Since $c_{n-1}=0$, the magnitudes of two positive numbers are added without any overflow.

$\Rightarrow$ no overflow.

Case III B: $a_{n-1}=0,\ b_{n-1}=1 \Rightarrow S = 1\,s_{n-2} \cdots s_0$

$$\Rightarrow A+B = -2^{n-1} + \sum_{i=0}^{n-2} (a_i + b_i)\cdot 2^i$$

$\Rightarrow$ Since $c_{n-1}=0$, there is no overflow in the lower $n-1$ bits of addition.

$$\Rightarrow s_{n-2}\,s_{n-3} \cdots s_0 = \sum_{i=0}^{n-2} (a_i + b_i)\cdot 2^i$$

$$\Rightarrow A+B = -2^{n-1} + s_{n-2} \cdots s_0 = 1\,s_{n-2} \cdots s_0$$

$\Rightarrow$ no overflow

Case IV : $c_{n-1} = c_n = 1$

$\Rightarrow$ At most one of a_{n-1} and b_{n-1} is 0.

Case IV A : $a_{n-1} = 1,\ b_{n-1} = 0 \Rightarrow S = 0\, s_{n-2} s_{n-3} \cdots s_0$

$$\Rightarrow A+B = -2^{n-1} + \sum_{i=0}^{n-2} (a_i + b_i)\, 2^i$$

$\Rightarrow$ Since $c_{n-1} = 1$, the result of adding the lower $n-1$ bits is $1\, s_{n-2} s_{n-3} \cdots s_0$ and this is treated as a positive number.

$$\Rightarrow \sum_{i=0}^{n-2} (a_i + b_i)\, 2^i = 2^{n-1} + \sum_{i=0}^{n-2} 2^i s_i$$

$$\Rightarrow A+B = -2^{n-1} + 2^{n-1} + \sum_{i=0}^{n-2} 2^i s_i$$

$$= \sum_{i=0}^{n-2} 2^i s_i$$

$$= 0\, s_{n-2} s_{n-3} \cdots s_0$$

$\Rightarrow$ no overflow

Case IV B : $a_{n-1} = b_{n-1} = 1 \Rightarrow S = 1\, s_{n-2} s_{n-3} \cdots s_0$

$$\Rightarrow A+B = -2^n + \sum_{i=0}^{n-2} (a_i + b_i)\, 2^i$$

$\Rightarrow$ Since $c_{n-1} = 1$, the result of adding the lower $n-1$ bits is $1\, s_{n-2} s_{n-3} \cdots s_0$ and this is treated as a positive number.

$$\Rightarrow \sum_{i=0}^{n-2} (a_i + b_i)\, 2^i = 2^{n-1} + \sum_{i=0}^{n-2} s_i\, 2^i$$

$$\Rightarrow A+B = -2^n + 2^{n-1} + \sum_{i=0}^{n-2} s_i\, 2^i$$

$$= -2^{n-1} + \sum_{i=0}^{n-2} s_i\, 2^i = 1\, s_{n-2} \cdots s_0$$

$\Rightarrow$ no overflow

1. Convert 6·75 to binary.

$6 = 110_2$

$0·75 \times 2 = 1·50$

$0·50 \times 2 = 1·00$

$6·75 = 110.11 = 1·1011 \times 2^2$

2. Convert 15·8125 to binary.

$15 = 1111_2$

$0·8125 \times 2 = 1·6250$

$0·625 \times 2 = 1·250$

$0·25 \times 2 = 0·50$

$0·50 \times 2 = 1·00$

$15·8125 = 1111.1101 = 1·1111101 \times 2^3$

3. Convert 1111·1101 to decimal.

$1111 = 15$

$0·1101 = 1 \times 2^{-1} + 1 \times 2^{-2} + 0 \times 2^{-3} + 1 \times 2^{-4}$

$= \frac{1}{2} + \frac{1}{4} + \frac{1}{16}$

$= 0·8125$

General formula: $(a_{n-1} a_{n-2} \cdots a_0 \cdot b_1 b_2 \cdots b_{k-1})_2$

$$= \sum_{i=0}^{n-1} a_i 2^i + \sum_{i=1}^{k-1} b_i 2^{-i}$$

Round to nearest:

Example in decimal : $2.378 \longrightarrow 2.38$

$2.374 \longrightarrow 2.37$

$2.375 \longrightarrow 2.38$ (nearest even)

In binary, it is always halfway rounding.

Example: 1.101 is to be rounded to two binary point places.

Example: $1.\underbrace{111\ldots1}_{24}$ is to be rounded to 23 mantissa bits.

The nearest neighbours : (i) $1.1111\ldots1$

$+\ 0.000\ldots1$

answer $\Longrightarrow$ $10.000\ldots00$ ← even

(ii) $1.1111\ldots1$

$-\ 0.000\ldots1$

$1.1111\ldots10$ ← not even

Example: $1.\underbrace{111\ldots1}_{22}01$ is to be rounded to 23 mantissa bits.

The nearest neighbours: (i) $1.111\ldots101$

$+\ 0.000\ldots001$

$1.111\ldots110$ ← not even

(ii) $1.111\ldots101$

$-\ 0.000\ldots001$

answer $\Longrightarrow$ $1.111\ldots100$ ← even

Round toward zero : +ve numbers decrease after rounding
-ve numbers increase after rounding

Round toward $+\infty$: all numbers increase after rounding

Round toward $-\infty$: all numbers decrease after rounding