

2/11/23

Computer Organization and Architecture

Unit-02

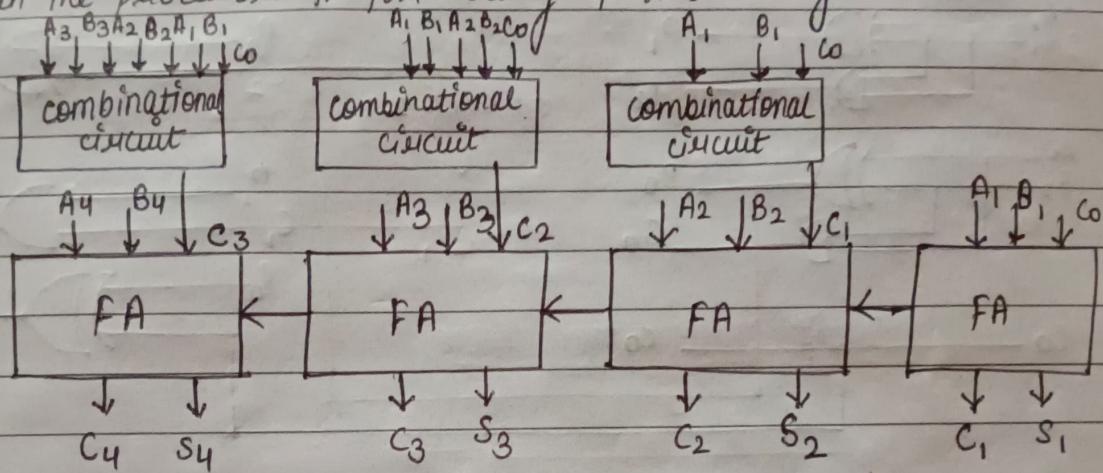
Arithmetic logic shift unit \rightarrow ① Responsible for execution.

Design of 4-bit ahead adder :-

- ② It contains arithmetic, logic & shift circuit.

\rightarrow It is used to speed up the delay

in the processor for adding of two binary numbers.

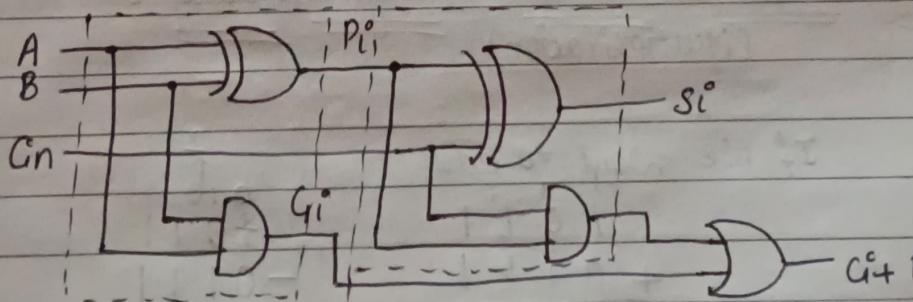


\rightarrow Full adder can be implemented using 2-Half adder and a OR-gate.

\rightarrow Idea is to look at the lower order bits to see if a higher order carry is to be generated, it uses two functions.

\rightarrow carry generate P_i^c and carry propagate G_i^c .

$$P_i^c = A_i \oplus B_i, \quad G_i^c = A_i \cdot B_i, \quad S_i^c = P_i^c \oplus C_i, \quad C_{i+1} = G_i^c + P_i^c C_i$$



$$C_{i+1} = G_i^c + P_i^c C_i \quad (G_0 + P_0 C_0)$$

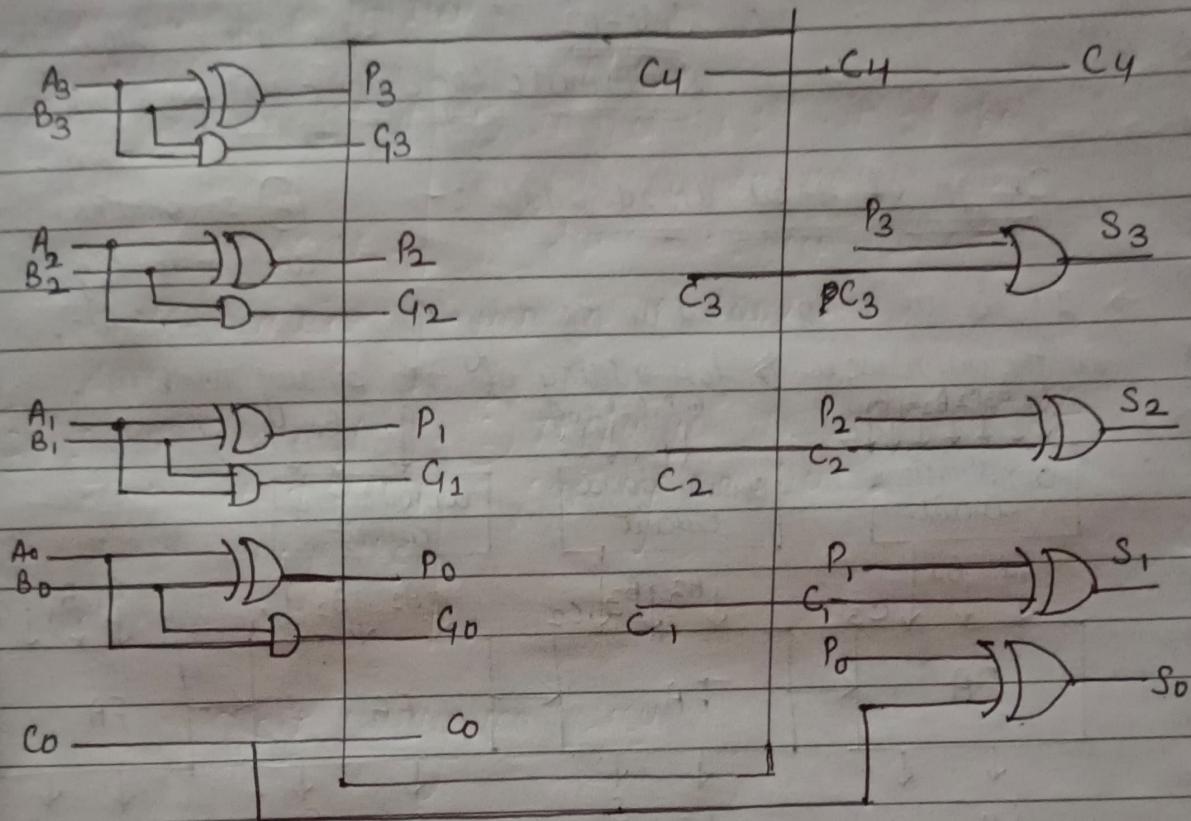
$$i=0, \quad C_1 = G_0 + P_0 C_0, \quad i=1 \Rightarrow C_2 = G_1 + P_1 C_1$$

$$C_3 = G_2 + P_2 C_2 \Rightarrow C_3 = G_2 + P_2 (G_1 + P_1 G_0 + P_0 G_0)$$

$\left\{ \begin{array}{l} P = \text{propagation} \\ g_i = \text{Generation} \end{array} \right\}$

Date.....

→ carry propagation problem has been resolved through carry look ahead adder.



Block Diagram

Disadvantage : Hardware is required to decrease the propagation delay.

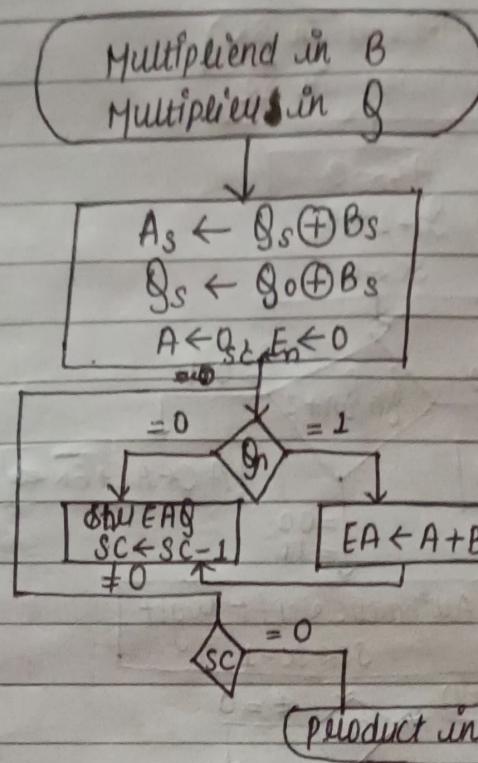
* For signed ~~complement~~ complement - magnitude data :-
(Multiplication)

If we want to add

$$\begin{array}{r}
 1 \ 0 \ 1 \ 1 \\
 + 0 \ 1 \ 1 \ 1 \\
 \hline
 1 \ 0 \ 0 \ 1 \ 0
 \end{array}$$

↓
Carry

$$\begin{array}{r}
 0 \ 1 \ 0 \\
 + 0 \ 1 \ 1 \\
 \hline
 - 0 \ 1 \ 1 \ 1
 \end{array}
 \quad \begin{array}{r}
 0 \ 1 \ 0 \ 0 \\
 \hline
 0 \ 1 \ 0 \ 0
 \end{array}$$

Algorithm :-Multiply operation

$$11 \times 9 = 99$$

11 - Multiplicand = 1011, 9 - Multiplier = 1001

E	A	Q	SC
0	0000	1001	4 = 100
Q _{n=1} , add B	+ 1011	1011	

shl EAQ

0	0101	1100	3 = 011
Q _{n=0} ,	0010	1110	2 = 010

shl EAQ

0	0001	0111	1 = 001
Q _{n=0}	+ 1011		

shl EAQ

0	1100		
Q _{n=1}			

add B

0	0110	0011	0 = 000
shl EAQ			

$$\text{Product} = A \bar{Q} = 01100011 \quad 1's \text{ complement}$$

$$= 64 + 32 + 2 + 1$$

$$= \underline{\underline{99}} \text{ down}$$

$$= 10011101$$

$$= 76543210$$

$$= 16 + 8 + 4 + 1$$

Spiral

The hardware for multiplication consist of register A & B. The multiplier is stored in the Q register & its sign in Q_3 . The sequence counter SC is initially set to a number equal to the no. of bits in the multiplier, the counter is decremented by 1 after forming each partial product. When the content of the counter reaches 0, the product is formed and process stops. Initially, the multiplicand is in register B and multiplier in Q. The sum of A & B forms a partial product which is transferred to EA register. Both partial product & multiplier are shifted to the right. This shift will be denoted by the statement "SHREAG" to designate the right shift. The least significant bit (LSB) of A is shifted into the MSB of Q.

The bit from E is shifted into MSB of A and 0 is shifted into E. After the shift, 1 bit of partial product is shifted into Q, pushing the multiplier bits 1 position to the right, in this manner the eight most flip flop in register Q designated by Q_n will hold the bit of multiplier which must be inspected next.

~~Handwritten algorithm :-~~

Initially, the multiplicand in B and multiplier in Q, their corresponding signs are in BS & QS respectively. The size are compared, & both A & Q are set to correspond to the sign of the product since a double length product will be stored in register A & Q.

Register A & E are cleared and the sequence counter SC is set to a number equal to the no. of bits of multiplier.

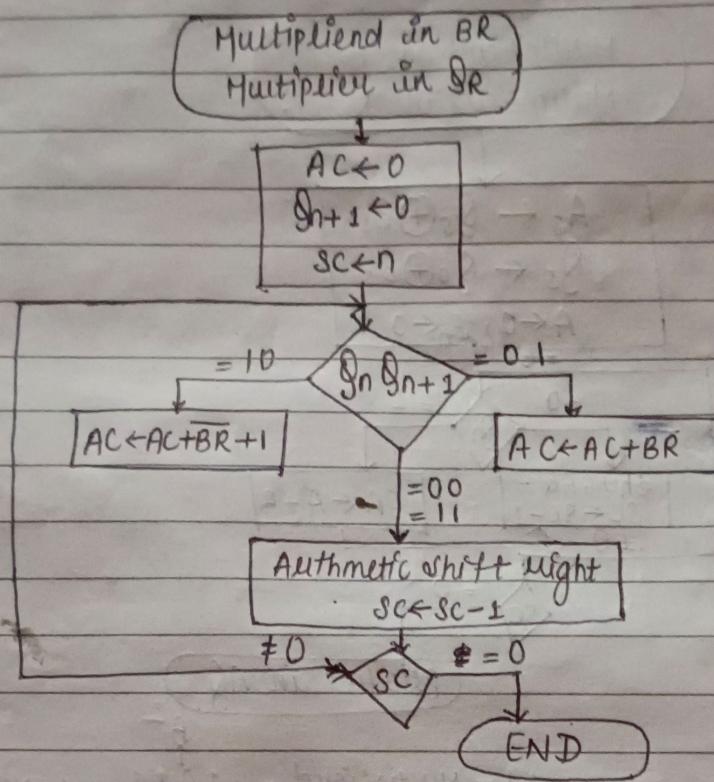
→ After the initialization, the lower product bit of the multiplier in Qn is tested, if it is 1, the multiplicand in B is added to the present partial product A, if it is 0, nothing is done, register EAQ is then shifted once to the right to form the new partial product. The SC is decremented by 1, and its new value checked. If it is not 0 then process is repeated and new partial product is formed.

→ The process stops when $SC=0$, note that the partial product formed in A is shifted into Q, 1 bit at a time & eventually replaces the multiplier, the product is available in both A & Q. A holding the MSB's & Q holding LSB's.

Date.....

Booth multiplication signed 2's complement :-

Multiplication



Ques $12 \times (-6) \Rightarrow 12 = \text{Multipliend} = BR = 01100$ (2's complement)
 $\overline{BR} + 1 = 10100$

$-6 = \text{Multiplier} = Q_R \Rightarrow 00110 \rightarrow 11010 = Q_R$
 2's complement of Q_R

$Q_n Q_{n+1}$	AC	QR	Q_{n+1}	Date.....
0 0	00000	1010	0	SC
ashu	00000	01101	0	$5 = 101$
1 0	00000			$4 = 100$
Sub BR	+ 10100			
	10100			
ashu	11010	00110	1	$3 = 011$
0 1	+ 01100			
Add BR	① 00110			
	↓ neglect			
ashu	00011	00011	0	$2 = 010$
1 0	+ 10100			
Sub BR	10111			
ashu	11011	10001	1	$1 = 001$
1 1	11101	11000	1	$0 = 000$
ashu				

Product = 1110111000 } 2's complement
 = 0001001000 ↙
 9876543210

= $64 + 8 = \boxed{72}$ ans

→ Booth algorithm give a process for multiplying binary integers in signed 2's complement representation, as in all multiplication scheme booth algorithm requires examination of multiplier bits & shifting of the partial product (P) to the shifting the multiplicand may be added to the partial product, subtract from the partial product or left unchanged according to the following rules :-

1. The multiplicand is sub. from the partial product upon encountering the first least significant one in strings of one's in the multiplier.
2. The multiplicand is added to the partial product upon encountering the first 0 in the string of zero's in the multiplier.
3. The partial product does not change when the multiplier bits are identical ~~are~~ to the previous multiplier bits.

Algorithm:-

The HW representation of booth algorithm require the register configuration similar to

Date.....

Multiplication of except that the signed bits are not separated from the rest of the registers. On designated the LSB of the ~~top~~ Multiplier in the register QR. An extra flip flop Q_{n+1} is appended to the QR to facilitate a double bit inspection, of the ~~excess~~ ^{multiplier} Register AC and appended bit Q_{n+1} are initially cleared to 0 & SC is set to a no. $n =$ to the no. of bits in the multiplier. The 2 bits of the multiplier Q_n & Q_{n+1} are inserted.

If the 2 bits are equal to 10, it means that the 1st 1 in a string of 1's has been encountered this requires a subtracting of multiplicand from the partial product of AC if the 2 bit are equal to 01, it means that the 1st 0 in a string of 0's has been encountered, this required the add of multiplicand to the partial product of AC, where the 2 bit are equal, the partial product has not changed and over flow cannot occur because the addition & subtraction of the multiplicand follows each other. The next step is to shift right the partial product & the multiplier (including the bit to $n+1$) this is a arithmetical shift right operation which shifts AC & B4 to the right and leave the signed bit in AC unchanged. The SC is implemented & the computational rule is repeated n-times.

Arithmetic multiplier :- Less complex and faster than booth and signed multiplier.

Arithmetic shift right :- 10110

110110 → rejected

Assignment shift left :- 10110

1 ← 0110 0 ← inserted
discarded

PROBLE

Spiral

Date.....

→ Home shift operations are used in Booth multiplication & signed multiplication. For better hardware management we use away multiplier.

ashu

① In ashu the 1st bit of the number remains the same.
i.e. MSB remains same.

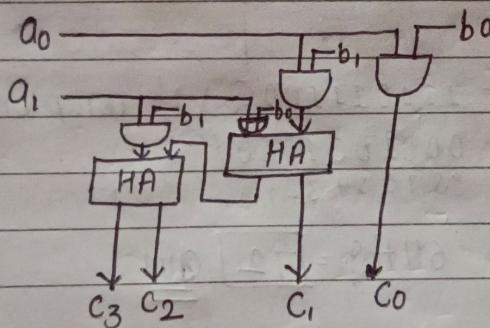
shu

① In this the 1st bit of the number does not remains the same i.e. MSB does not remain the same.

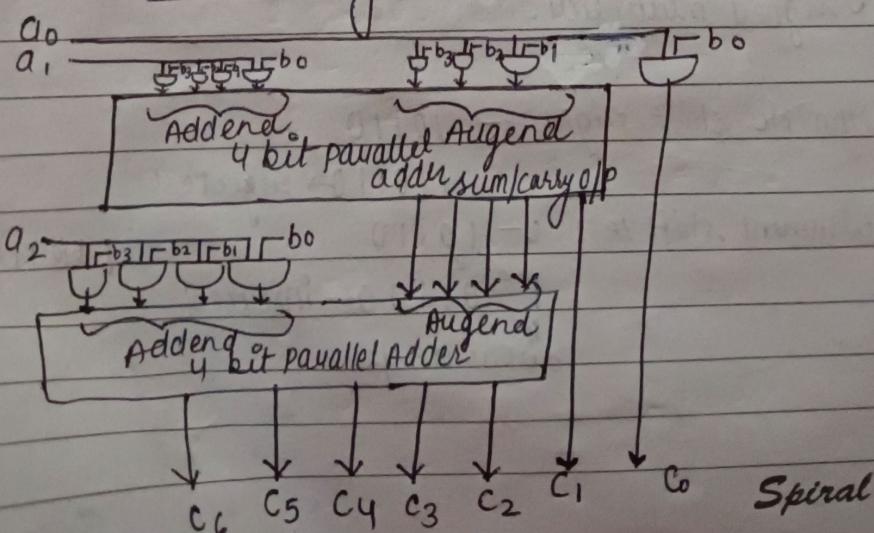
② In this carry is neglected and doesn't change the value of E.

② It changes the value of E if carry appears.

2 bit away multiplier



4 bit away multiplier



Spiral

Date.....

2 by 2 bit

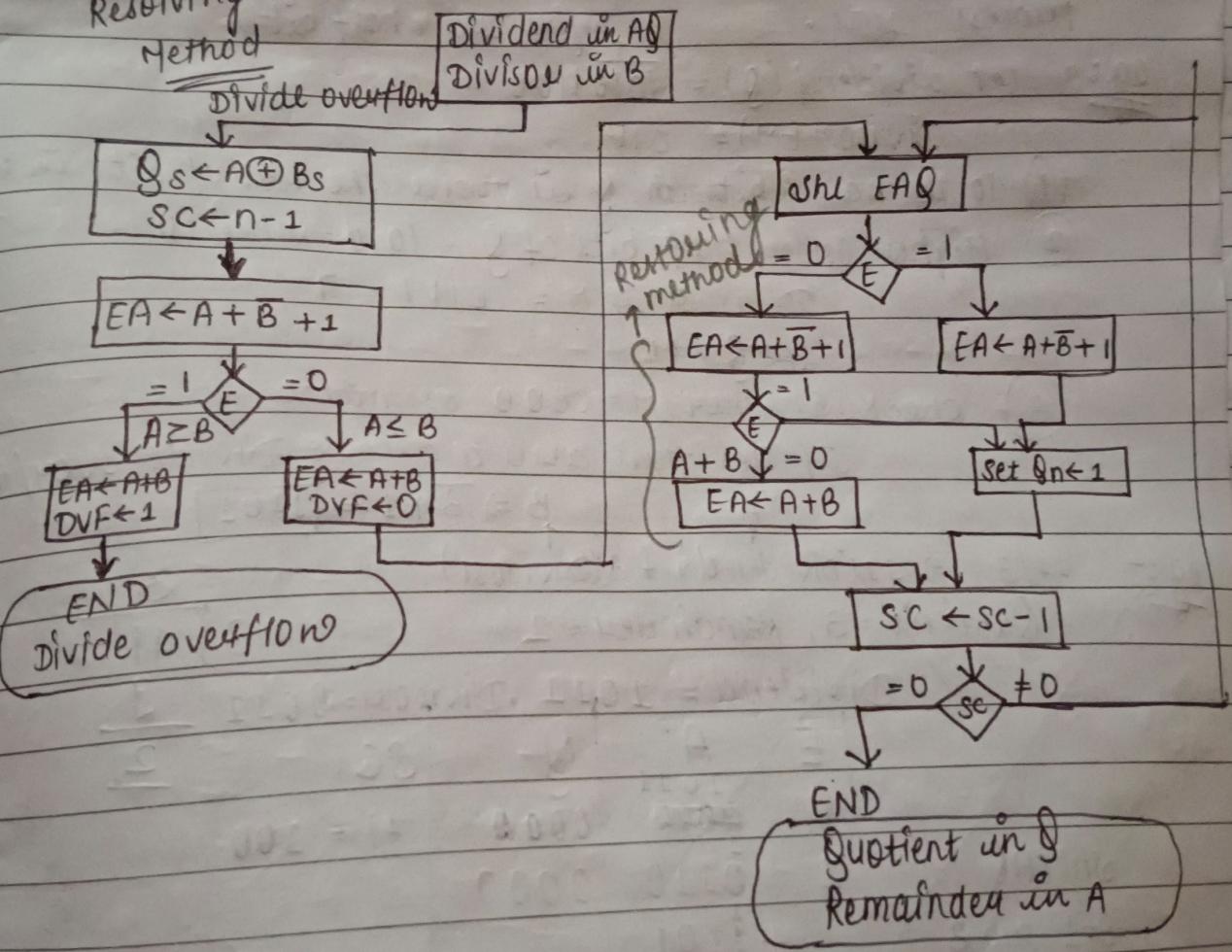
$$\begin{array}{r}
 a_1 \quad a_0 \rightarrow \text{Multiplicand} \\
 b_1 \quad b_0 \rightarrow \text{Multiplier} \\
 a_1 b_0 \quad a_0 b_0 \\
 + a_1 b_1 \quad a_0 b_1 \quad \longrightarrow \text{Shift 1 bit} \\
 c_2 \quad c_1 \quad c_0
 \end{array}$$

Division algorithm

Resolving
Method

Divide overflow

Dividend in A
Divisor in B



Divide overflow condition occurs when :-

- The no. of bits in dividend is twice as the no. of divisor.
- The higher order half bits of the dividend has a value more than or equal to the value of the total bit of the divisor.

(iii) The divide overflow also check that whether the divisor is zero or not.

(iv) All these conditions are checked before the division process starts.

Eg: Let dividend (Q) = $\overline{10101100}$

Divisor (M) = 0111

- so the no. of bits in Q is twice the no. of bits in M .
- Higher order half bits of Q = 1010 = value 10

Division of M = $0111 = 7$

so $Q > M$

③ Check whether $M = 0000$ or not

If yes then also indicate Divide overflow

$$\begin{array}{r} A \\ \uparrow \\ B = \overline{B} + 1 = \boxed{1101} \end{array}$$

Eg:- $11/3 \Rightarrow 11$ (Dividend) & 3 (Divisor)

~~Restoring Method~~

$$\text{Quotient} = \overset{-001}{3}, \text{Remainder} = \overset{-0010}{2} \quad 3 \overline{) 11}$$

~~Dividend = 1011, Divisor = 0011~~

$$\frac{3}{2}$$

~~| E | A | Q | SC |
|---|---------------------|------|-----------|
| 0 | 0000 | 0000 | $4 = 100$ |
| 1 | 0110 | 0000 | |
| | $+ \overline{1101}$ | | |
| 1 | 0011 | | |~~

~~$E = 1, \text{set } Q \leftarrow 1$~~

~~$E = 0, \text{sub } B$~~

~~$E = 1, \text{set } Q \leftarrow 1$~~

~~$E = 0, \text{sub } B$~~

~~$E = 1, \text{set } Q \leftarrow 1$~~

Eg: $11/3 = \text{Dividend} = 11 + \text{Division} = 3$

Quotient = 3, Remainder = 2

Dividend = A = 01011, Division = B = 0011
 $\overline{B} + 1 = 1101$

	E	A	Q	SC
Shl EA Q	0	0000	1011	4 = 100
Sub B	0	0001	0110	
E=0	0	<u>1101</u>		
E=0, add B	+	<u>0011</u>		
	1	0001	0110	3 = 011
Shl EA Q	0	0010	1100	
E=0, Sub B	+	<u>1101</u>		
E=0, add B	0	<u>1111</u>		
	1	0010	1100	2 = 010
Shl EA Q	0	0101	1000	
E=0, Sub B	+	<u>1101</u>		
E=1, set Qn-1	1	0010	1001	1 = 001
Shl EA Q	0	<u>0101</u>		
E=0, sub B	+	<u>1101</u>		
E=1, set Qn-1	1	0010	0011	0 = 000

Quotient in Q = 0011 = 3

Remainder in A = 0010 = 2

Date.....

The dividend is in A_Q and divisor is in B₀. The sign of the dividend is transferred into Q₃ to be the part of the quotient. The constant is set into the sequence counter so to specify the number of bits in the ~~quotient~~. As in multiplication we assume that operands are transferred to registers from memory ~~one~~ until ~~one~~ that has words of n bits. ~~and~~ Since an operand must be stored with its sign. 1 bit of the word will be occupied by the sign and magnitude will be consists of n-1 bits. If divide overflow condition is tested by subtracting the divisor in B from half of the bits of the dividend stored in A. If A ≥ B, the divide overflow flag (DVF) is set and the operation is terminated prematurely. If A < B & no overflow divide occurs so the value of the dividend is restored by adding B to A. The division of the magnitude starts by shifting the dividend in A_Q to the left with the high order bit shifted into E. If the bit shifted into E is 1 we know that EA > B bcoz EA consists of a 1 followed by n-1 bits while B consists of only n-1 bits. In this case, B must be subtracted from EA and 1 inserted into Q_n for the quotient bit. Since register A is missing the high order bit of the dividend which is in E() its value is $EA - 2^{n-1}$ adding to this value the 2's complement of B result in $(EA - 2^{n-1}) + (2^{n-1} - B) = EA - B$. The carry from this ~~addition~~ addition is not transferred to E. If g we want to E to remain as 1 the shift left operation inserts a 0 into E the divisor is subtracted by adding its 2's complement value and carry is transferred into E. If E=1 it signifies that

ALB and original number is restored by adding B to A in the later case we leave 0 in Qn. This process is repeated again with register & holding the partial remainder after $n-1$ time. The Quotient ~~is formed~~ magnitude is form in register Q & remainder is found ~~in~~ in register the Quotient sign is in Q0 & the sign of sum is in A0, is the same as the original sign of the dividend.

Types of division :-

- ① Restoring Method :- In this method the partial remainder is restored by adding the divisor to the -ve differences.
- ② comparison Method :- In this method A & B are compared prior to the subtraction operation. If $A \geq B$, B is subtracted from A so if $A < B$ nothing is done. The partial remainder is shifted left and the numbers are compared again. The comparison is determined prior to the subtraction by inspecting the ~~END~~ carry out of the parallel adder period (to its transfer to register E).
- ③ Non restoring Method :- In this method B is not added if the difference is -ve but instead the -ve difference is shifted left and then B is added. We note that the operation perform a $A - B + B$ i.e. B subtracted and then added to restore A. The next time around the loop this number is shifted left (or multiplied by 2). The next time around the loop we subtract the value again this gives $2(A - B) - B = 2A - B$. This result is obtained in the non-restoring method by leaving $A - B$ as it. The next time around the loop the number is shifted left & B added to give $A - B + B$ i.e. equal to A thus

Writ



Date.....

is same as before. Thus in the non-restoring method B is subtracted if the previous value of Qn was 1 but B is added in previous value of Qn was 0 and no restoring of the partial remainder is required. This process save the steps to adding the divisor of A < B but it requires special control logic to remember the previous result. The first time the dividend is shifted, B must be subtracted. Also if the last bit of the quotient is 0, the partial remainder must be restored to obtain the correct final remainder.

→ Divide overflow condition occurs if the high order half bits of the dividend consists a number greater than or equal to the divisor. Another problem associated with division is the fact that a divide by 0 must be avoided. ~~Another problem associated~~ This occurs becoz ~~any~~ any dividend will be greater than or equal to an divisor which is equal to an 0.

* Floating point Arithmetic operations :-

① Addition / subtraction of floating point numbers :-

Algo :-

- ① Check for zero's, if either no. is zero, the result will be other number with appropriate sign.

mantissa

\uparrow exponent

$$\text{Let, } X = X_A * 2^{F_A}$$

$$Y = X_B * 2^{F_B}$$

During addition :- If $y=0$, $Z=x=X_A * 2^{F_A}$

$$\text{If } x=0, z=y = X_B * 2^{F_B}$$

During subtraction :- If $y=0$, $Z=x=X_A * 2^{F_A}$

$$\text{If } x=0, z=-y = -X_B * 2^{F_B}$$

- ② Align the mantissa :- If exponent are equal, perform the arithmetic operations ; else shift the mantissa to the right with smaller exponent until its exponent becomes equals to the larger exponent.

Let $x = 1.10 * 2^{+3}$ (larger exponent)

$y = 1.01 * 2^{+1}$ (smaller exponent)

$$Z = x+y$$

Shift the mantissa of y to the right, until its exponent is equal to the exponent of x .

$$y = 0.101 * 2^{+2}$$

$$y = 0.0101 * 2^{+3}$$

$$x = 1.10 * 2^{+3}, y = 0.0101 * 2^{+3}$$

- ③ Depending upon the operation & sign of the two mantissa perform addition / subtraction.

* During addition, if two significant are equal with opposite sign, the result will be zero.

Date.....

* There is also a possibility of significant (mantissa) overflow by 1 digit, then the mantissa of the result is shifted right and exponent is incremented.

Significant overflow :- occurs, when the addition of two significant (mantissa) of same sign may result in a carry at MSB of mantissa as we increment exponent, there is also a possibility of exponent overflow. The result will show "error".

① If

$$x = 1.01 * 2^3$$

$$y = -1.01 * 2^3$$

$$\text{then } z = x + y$$

$$0 * 2^3 = 0$$

②

$$\text{If } x = 1.01 * 2^3$$

$$y = \underline{1.10 * 2^3}$$

$$\textcircled{1} \underline{0.11 * 2^3}$$

↓ carry in MSB

$$\text{Correction} = \boxed{1.011 * 2^4}$$

④ Normalize the result :- If result is not normalized, then we shift mantissa left and decrement the exponent until value "left of binary point is 1". As we are decreasing the exponent, so exponent can also be underflow then again error is reported.

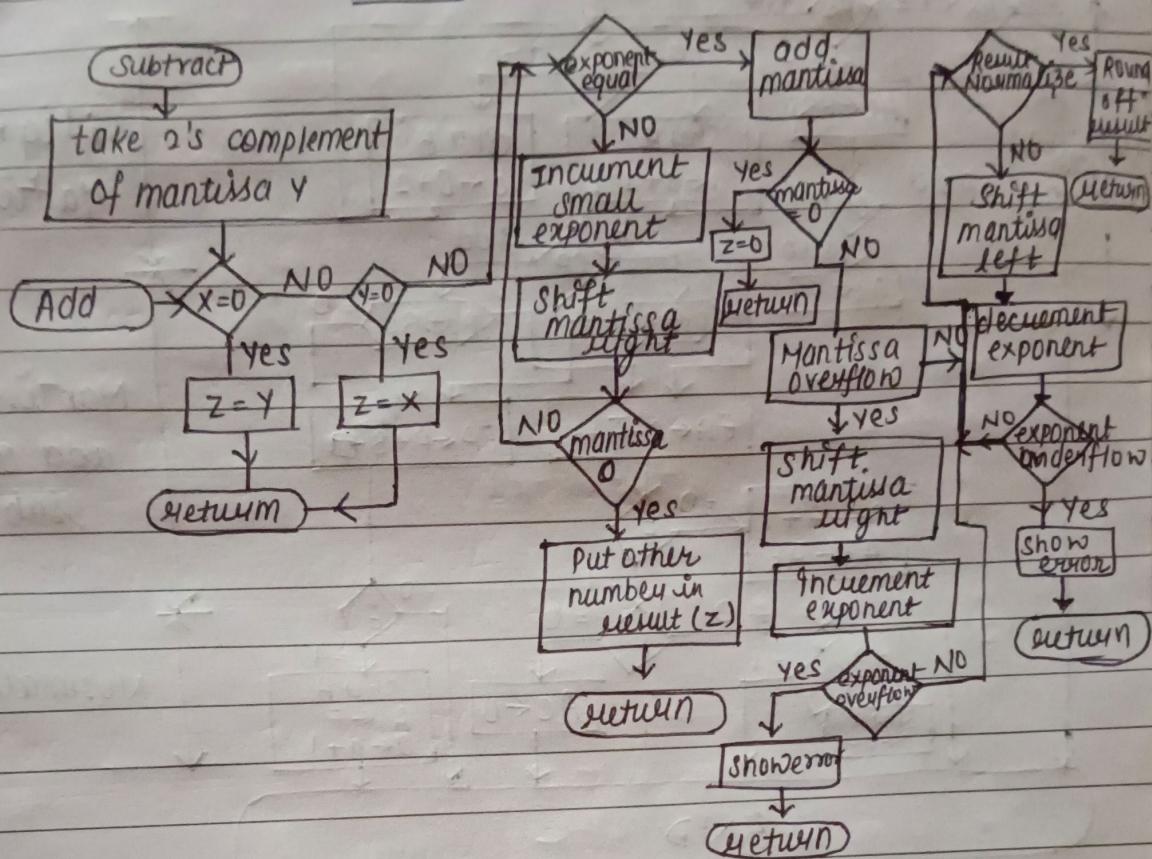
$$\text{If } x = 1.01 * 2^3$$

$$-1.00 * 2^3$$

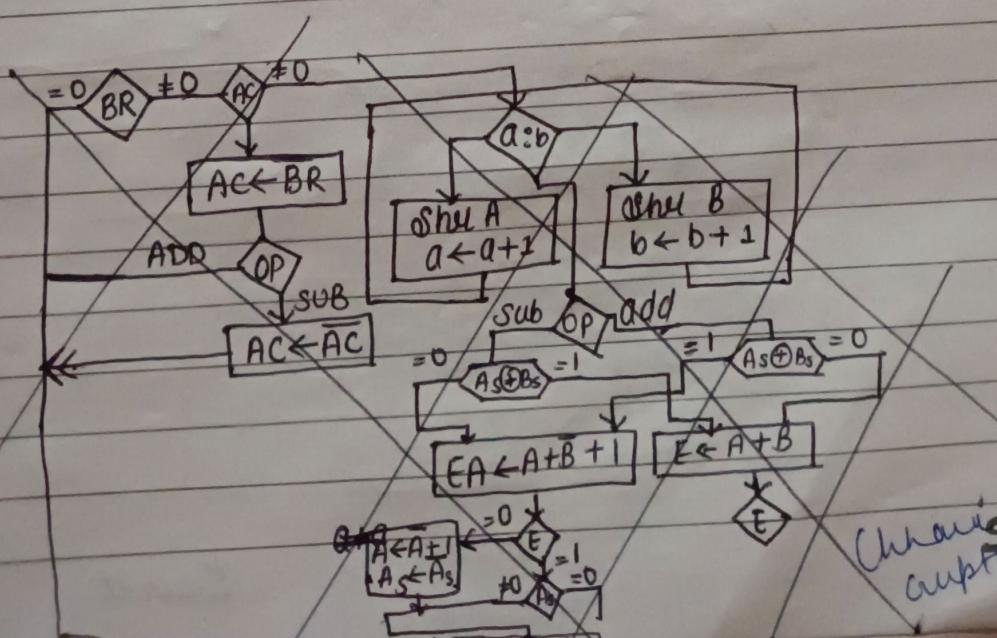
$$\underline{0.01 * 2^3} \quad (\text{NOT normalized})$$

$$1.0 * 2^1 \quad (\text{Normalized})$$

Flowchart

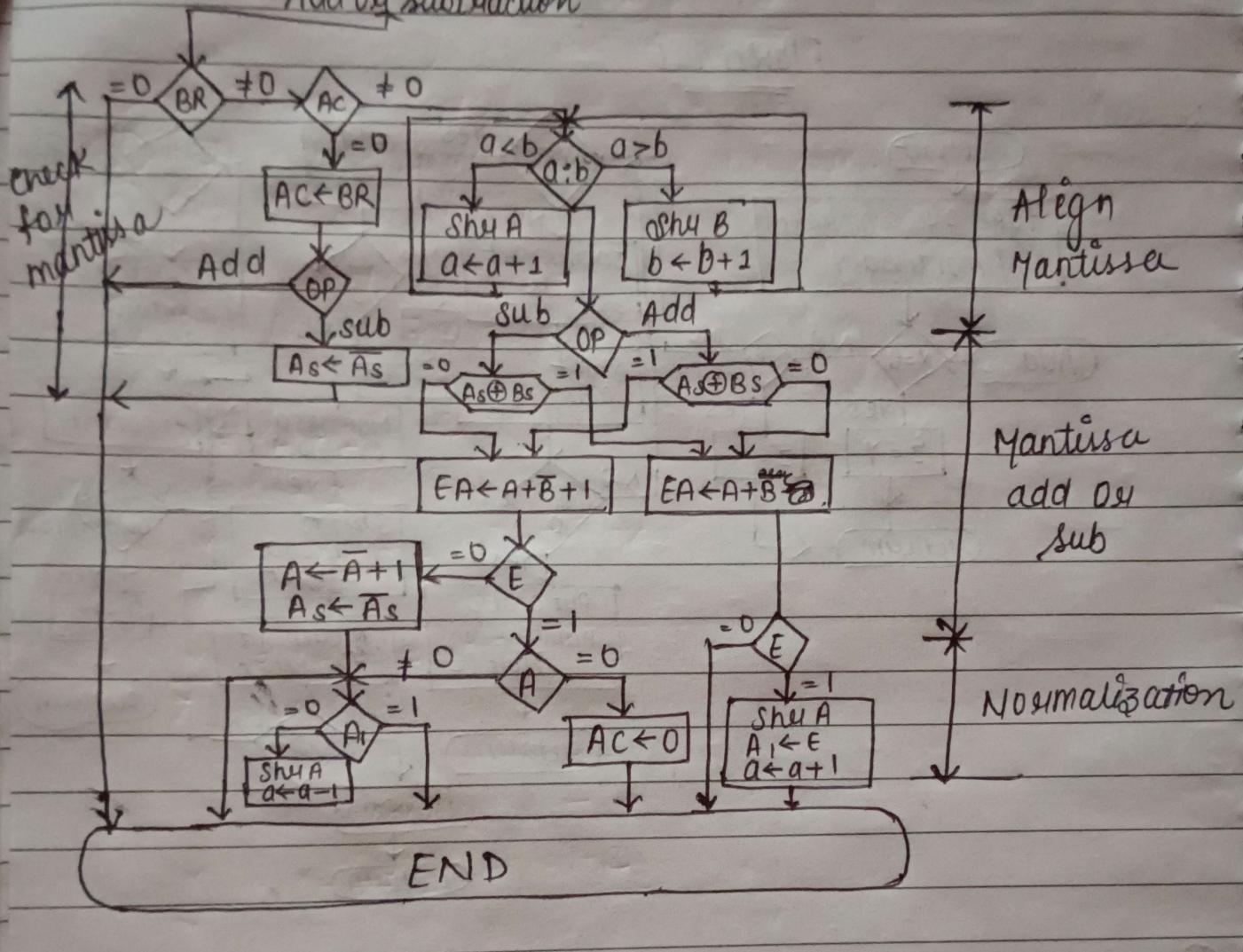


- ① Significant overflow :- When two similar sign significant (mantissa) are added, the sum may result in a carry out of the MSB. This is called significant overflow.
 - ② Significant underflow :- If the resulting value has 0 in the MSB.



Date.....

Add or subtraction



Add or sub for floating point numbers

Floating point multiplication

- ① Check for zeros.
- ② Add the exponents.
- ③ Multiply the mantissa.
- ④ Normalize the product.

$$X = X_A * 2^{E_A + e} \xrightarrow{\text{biased exponent}}$$

$$Y = X_B * 2^{E_B + e}$$

$$X * Y = (X_A * 2^{E_A + e}) (X_B * 2^{E_B + e})$$

$$Z = (X_A * X_B) (2^{E_A + E_B + 2e})$$

$$Z = (X_A * X_B) (2^{E_A + E_B + 2e - e} \text{ (as exponent is added twice)})$$

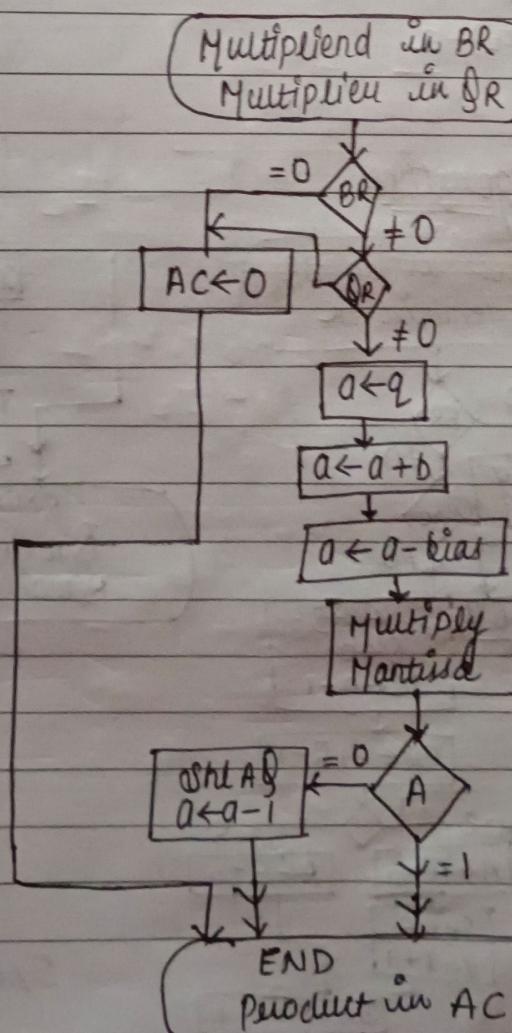
$$Z = (X_A * X_B) (2^{E_A + E_B + e})$$

$$Z = X * Y$$

$$\begin{matrix} \swarrow 0 \\ Z = 0 \\ \downarrow = 0 \end{matrix}$$

$$Z = 0$$

Algorithm



Floating point four division

- ① Check for zeros.
 - ② Initialize registers & evaluate sign.
 - ③ Align the dividend.
 - ④ Subtract the exponent
 - ⑤ Divide the mantissa.

$$x = x_A * 2^{E_A + e}$$

$$Y = X_B * 2^{EB} + e$$

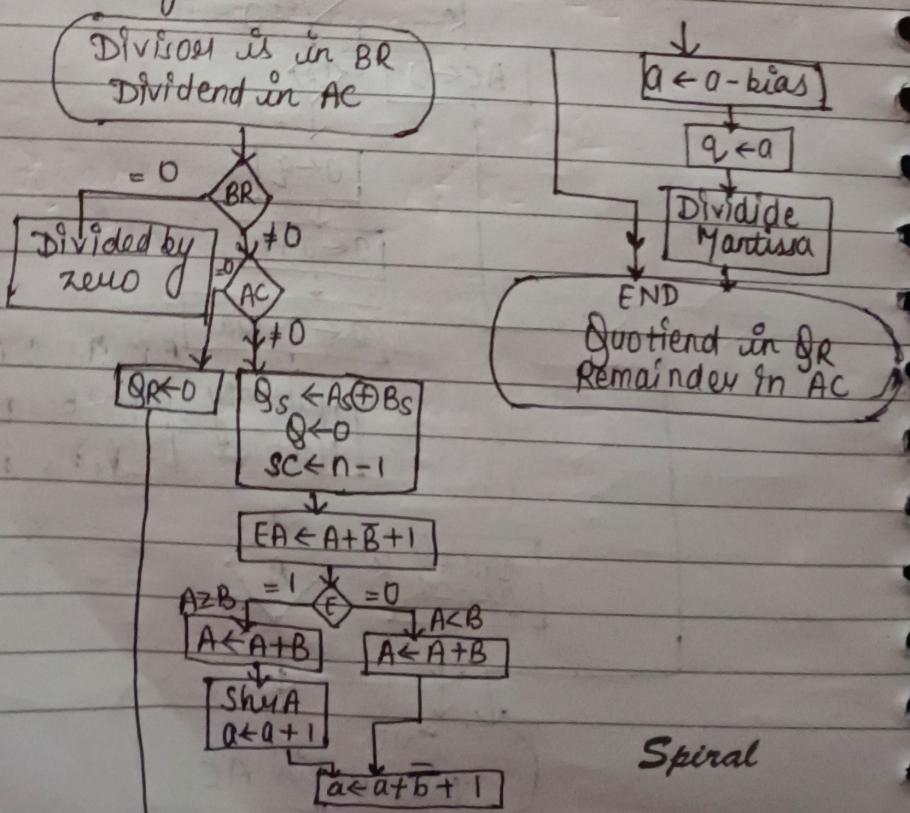
$$\frac{X}{Y} = \frac{XA}{XB} * 2^{(EA+g-EB-d)}$$

$$\frac{x}{Y} = \frac{x_A}{x_B} * 2^{(E_A - E_B)}$$

biasing exponent

$$\frac{x}{y} = \frac{x_A}{x_B} * 2^{(E_A - E_B) + e}$$

Algorithm

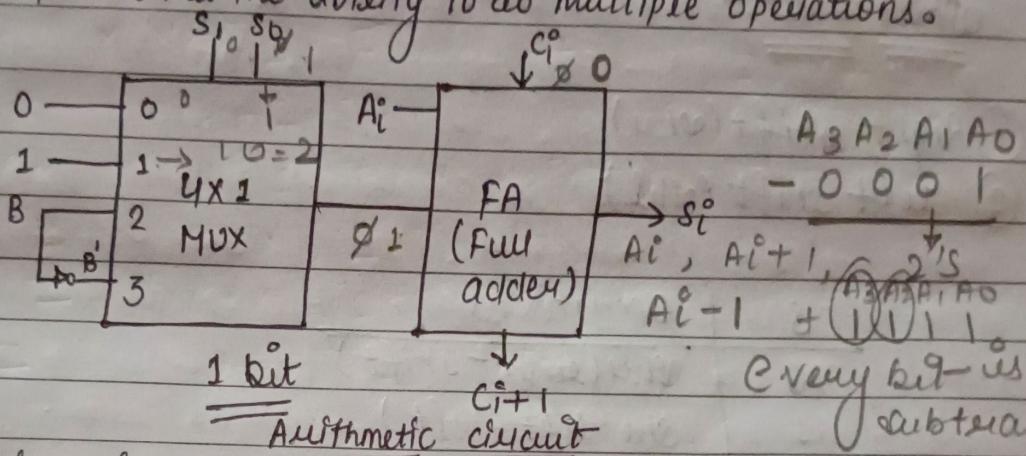


Arithmetical and Logic unit Design

$ALU = \text{Arithmetical unit} + \text{Logic unit} + \text{shift unit}$

Arithmetical circuit

→ Multiplexer has the ability to do multiple operations.



→ Selection Lines decides which opn will perform.

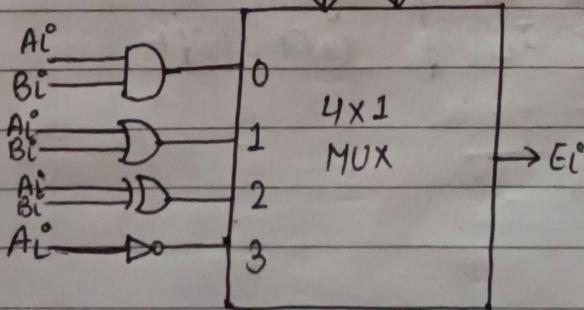
S_i	S_0	C_i^0	Operations
0	0	0	A_i^0 (Transfer)
1	0	0	A_i^{i+1} (Increment)
2	0	1	A_i^0 (Decrement)
3	0	1	A_i^0 (Transfer)
4	1	0	$A_i^0 + B_i^0$ (Add)
5	1	0	$A_i^0 + B_i^{i+1}$ (Add with carry)
6	1	1	$A_i^0 + \bar{B}_i^0$ (Subtract with borrow)
7	1	1	$A_i^0 + \bar{B}_i^{i+1}$ (Subtract)

Logic unit :-

S_1 S_0

At a time one opn has been performed

i.e. we use multiplexer.



1 bit

Logic circuit

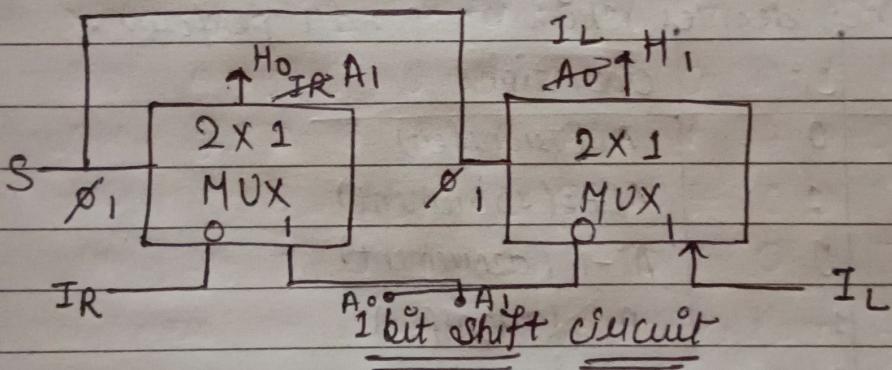
Spiral

Si	So	Operation (EP)
0	0	AND($A_i \cdot B_j$)
0	1	($A_i + B_j$) OR
1	0	($A_i \oplus B_j$) XOR
1	1	A_i' (NOT)

Shift unit :-

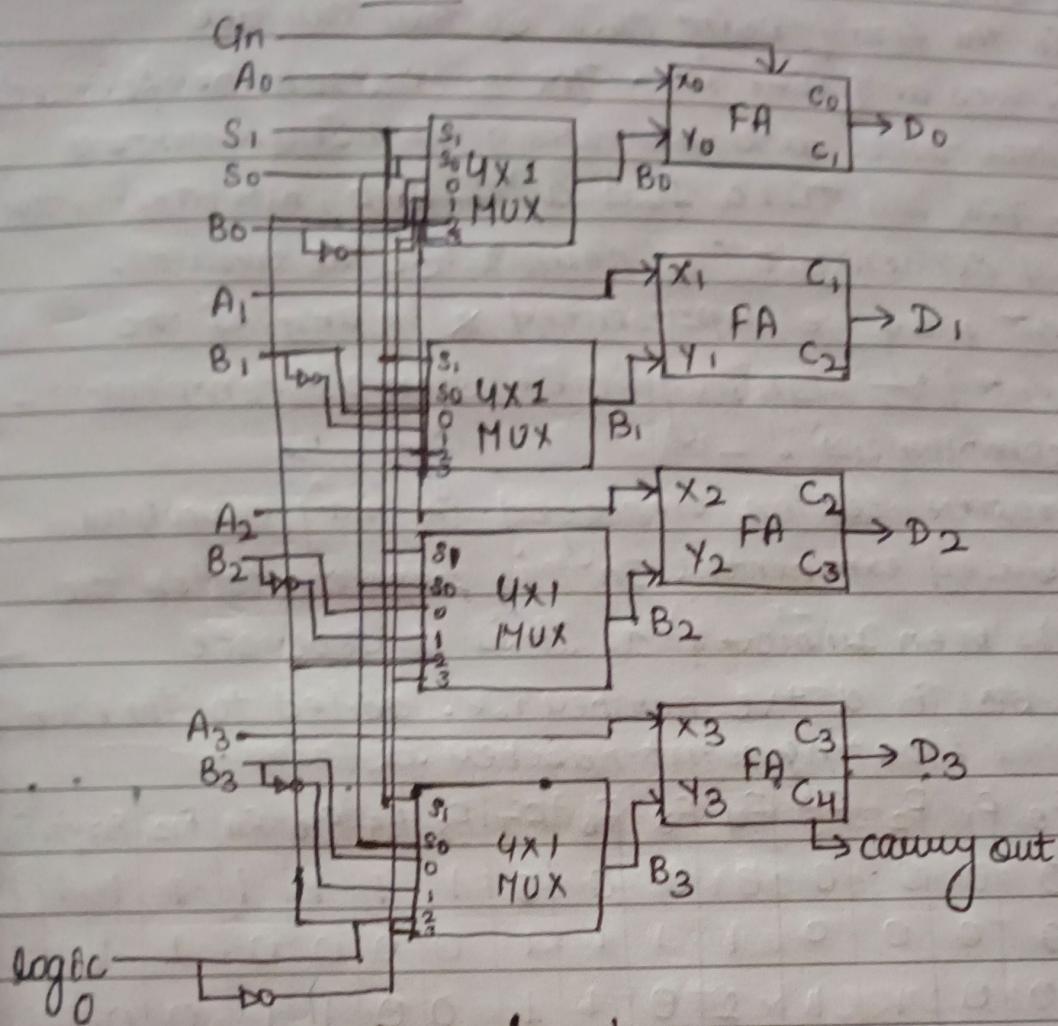
- ① Right shift (shru)
- ② Left shift (shl)

S	Operations
0	shru (right shift)
1	shl (left shift)



$$\begin{array}{c}
 \nearrow A_0 \swarrow A_1 \\
 \nearrow A_0 \quad \searrow A_1 \\
 \text{IR} \qquad \qquad \qquad A_0 \\
 \end{array}
 \qquad
 \begin{array}{c}
 \nearrow A_0 \swarrow A_1 \\
 \nearrow A_1 \quad \searrow I_L \\
 A_0 \qquad \qquad \qquad I_L
 \end{array}$$

S	H0	H1	Operation
0	IR	A0	Right shift
1	AI	IL	Left shift

ALU

logic 0

S_3	S_2	S_1	S_0	C_i	f_i (operation)
0	0	0	0	0	A_i^i (Transfer)
0	0	0	0	1	A_{i+1}^i (Increment)
0	0	0	1	0	A_{i-1}^i (Decrement)
0	0	0	1	1	A_i^i (Transfer)
0	0	1	0	0	$A_i^i + B_i^i$ (Add)
0	0	1	0	1	$A_i^i + B_i^i + 1$ (Add with carry)
0	0	1	1	0	$A_i^i + \overline{B_i^i}$ (Sub with borrow)
0	0	1	1	1	$A_i^i + \overline{B_i^i} + 1$ (Subtract)
0	1	0	0	X	AND (logic)
0	1	0	1	X	OR (logic)
0	1	1	0	X	XOR (logic)
0	1	0	1	X	NOT (logic)
1	0	X	X	X	Shift right
1	1	X	X	X	Shift left

Date.....

- The operation performed with the mantissa's are same as in the fixed point no.s so they can use some registers and circuits. The operation performed with the exponents are compared and increments of to align the mantissa for add, sub, multiplication & divide and decrement (to normalize the result). The exponent may be represented in any of the three represented as sign magnitude, sign 2's complement & sign 1's complement.
- Fourth representation employed in many computers known as biased exponents. In this representation the sign bit has been removed as it is a separate entity. The bias is a +ve number i.e. added to the each exponent as the floating point number is formed.

x	y	F ₀	F ₁	F ₂	F ₃	F ₄	F ₅	F ₆	F ₇	F ₈	F ₉	F ₁₀	F ₁₁	F ₁₂	F ₁₃	F ₁₄	F ₁₅
0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1
0	1	0	0	0	0	1	1	1	1	0	0	0	1	1	1	1	1
1	0	0	0	1	1	0	0	1	1	0	0	1	1	0	0	1	1
1	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1	0	1

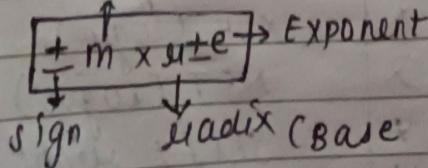
eg: consider an exponent that ranges from -50 to +49. Internally it is represented by +ve number (without a sign).

IEEE standard for floating point numbers

$$215 = \text{fixed point number} \Rightarrow 21.5 \times 10^1 \Rightarrow 2.15 \times 10^2$$

Floating point number contains two parts:

- (i) Mantissa: contains signed fixed point number
- (ii) Exponent: specify the position of decimal binary point.



examples :-

$$+215.37 = +0.21537 \times 10^{+3}$$

↓ ↓ ↓ → exponent
sign Mantissa Base

2	25
2	121
2	60
2	31

$$1000.110 = +0.1000110 \times 2^{+4}$$

↑ ↓ ↑ → exponent
sign Mantissa Base (Mantissa)

0.1101

$$-101011 = -0.101011 \times 10^{+6}$$

↓ ↑ → exponent
= sign Mantissa

These numbers are called floating point numbers because of the decimal (binary) point is fluctuating (floating).

Normalization :-

A floating point number is said to be normalized, when we force the integer part of its ~~magnitude~~ ^{Mantissa} to be 1, and allow its fraction part to anything.

e.g.:- $13.25 = 1101.01$

$$= 1.10101 \times 2^{+3} \quad (\text{Normalized})$$

↓ ↗
integer fraction
part part

$$= 1.ffff \times 2^{\text{exponent}}$$

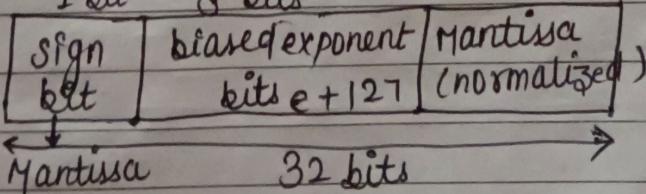
$$\begin{array}{r} 0.25 \\ \times 2 \\ \hline 0.50 \\ \times 2 \\ \hline 1.00 \\ \end{array}$$

$$0.25 \times 2 = 0.5$$

$$0.5 \times 2 = 1.0$$

IEEE standard :-

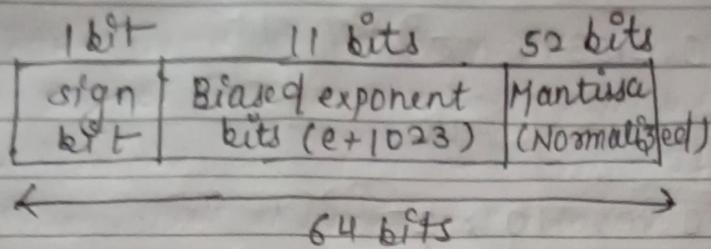
① single precision floating point standard :- / 32 bit floating point representation



positive $\rightarrow 0$
negative $\rightarrow 1$

Date.....

② Double precision floating point standard :-



Biased exponent for 32 bits

$$-(2^{8-1}) \text{ to } +(2^{8-1} - 1)$$

$$-127 \text{ to } +127$$

for 64 bits

$$-(2^{11-1} - 1) \text{ to } (2^{11-1} - 1)$$

Ques Represent 85.125 in IEEE 754 single precision floating point representation.

Ans 85.125 (decimal)

= 1010101.001 (binary)

so,

$$1.010101001 \times 2^6$$

$$\text{exponent} = e = 6, \text{ sign bit} = 0$$

$$\text{Biased exponent} = 6 + 127 = 133 = 10000101$$

$$\begin{matrix} \text{Mantissa} & = 010101001 \\ & (\text{Normalized}) \end{matrix}$$

1 bit	8 bit	23 bits
0	10000101	01010100 000000000