MODULE 4: ARITHMETIC

NUMBERS, ARITHMETIC OPERATIONS AND CHARACTERS

NUMBER REPRESENTATION

- Numbers can be represented in 3 formats:
  1) Sign and magnitude
  2) 1's complement
  3) 2's complement
- In all three formats, MSB=0 for +ve numbers & MSB=1 for -ve numbers.
- In **sign-and-magnitude system**,

  negative value is obtained by changing the MSB from 0 to 1 of the corresponding positive value.

  For ex, +5 is represented by 0101 &

  -5 is represented by 1101.
- In **1's complement system**,

  negative values are obtained by complementing each bit of the corresponding positive number.

  For ex, -5 is obtained by complementing each bit in 0101 to yield 1010.

(In other words, the operation of forming the 1's complement of a given number is equivalent to subtracting that number from 2^n-1).

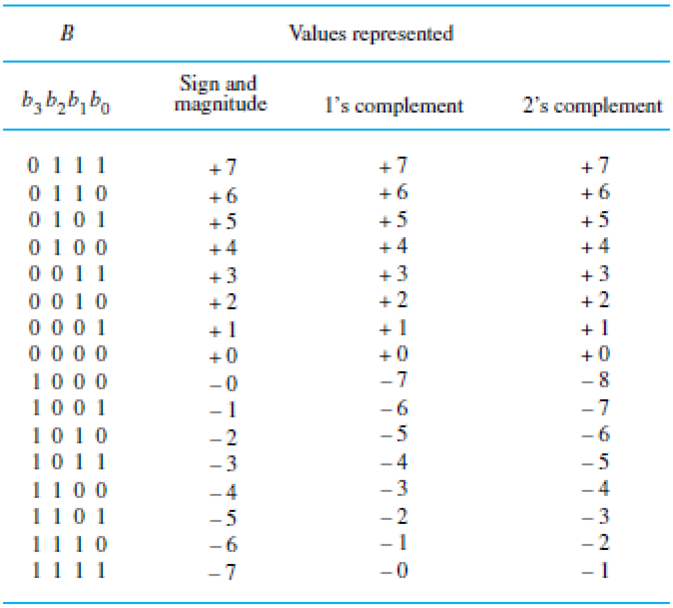
- In **2's complement system**,

  forming the 2's complement of a number is done by subtracting that number from 2^n.

  For ex, -5 is obtained by complementing each bit in 0101 & then adding 1 to yield 1011.

(In other words, the 2's complement of a number is obtained by adding 1 to the 1's complement of that number).

- 2's complement system yields the most efficient way to carry out addition/subtraction operations.



B	Values represented		
$b_3 b_2 b_1 b_0$	Sign and magnitude	1's complement	2's complement
0 1 1 1	+7	+7	+7
0 1 1 0	+6	+6	+6
0 1 0 1	+5	+5	+5
0 1 0 0	+4	+4	+4
0 0 1 1	+3	+3	+3
0 0 1 0	+2	+2	+2
0 0 0 1	+1	+1	+1
0 0 0 0	+0	+0	+0
1 0 0 0	−0	−7	−8
1 0 0 1	−1	−6	−7
1 0 1 0	−2	−5	−6
1 0 1 1	−3	−4	−5
1 1 0 0	−4	−3	−4
1 1 0 1	−5	−2	−3
1 1 1 0	−6	−1	−2
1 1 1 1	−7	−0	−1

Figure 1.3 Binary, signed-integer representations.

ADDITION OF POSITIVE NUMBERS

- Consider adding two 1-bit numbers.
- The sum of 1 & 1 requires the 2-bit vector 10 to represent the value 2. We say that sum is 0 and the carry-out is 1.

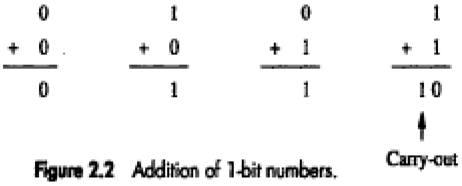


Figure 2.2 Addition of 1-bit numbers.

ADDITION & SUBTRACTION OF SIGNED NUMBERS

• Following are the two rules for addition and subtraction of n-bit signed numbers using the 2's complement representation system (Figure 1.6).

Rule 1:

➢ **To Add** two numbers, add their n-bits and ignore the carry-out signal from the MSB position.

➢ Result will be algebraically correct, if it lies in the range -2^{n-1} to $+2^{n-1}-1$.

Rule 2:

➢ **To Subtract** two numbers X and Y (that is to perform X-Y), take the 2's complement of Y and then add it to X as in rule 1.

➢ Result will be algebraically correct, if it lies in the range (2^{n-1}) to $+(2^{n-1}-1)$.

• When the result of an arithmetic operation is outside the representable-range, an arithmetic overflow is said to occur.

• To represent a signed in 2's complement form using a larger number of bits, repeat the sign bit as many times as needed to the left. This operation is called **sign extension**.

• In 1's complement representation, the result obtained after an addition operation is not always correct. The carry-out(c_n) cannot be ignored. If c_n=0, the result obtained is correct. If c_n=1, then a 1 must be added to the result to make it correct.

OVERFLOW IN INTEGER ARITHMETIC

• When result of an arithmetic operation is outside the representable-range, an **arithmetic overflow** is said to occur.

• For example: If we add two numbers +7 and +4, then the output sum S is 1011(←0111+0100), which is the code for -5, an incorrect result.

• An overflow occurs in following 2 cases

1) Overflow can occur only when adding two numbers that have the same sign.

2) The carry-out signal from the sign-bit position is not a sufficient indicator of overflow when adding signed numbers.

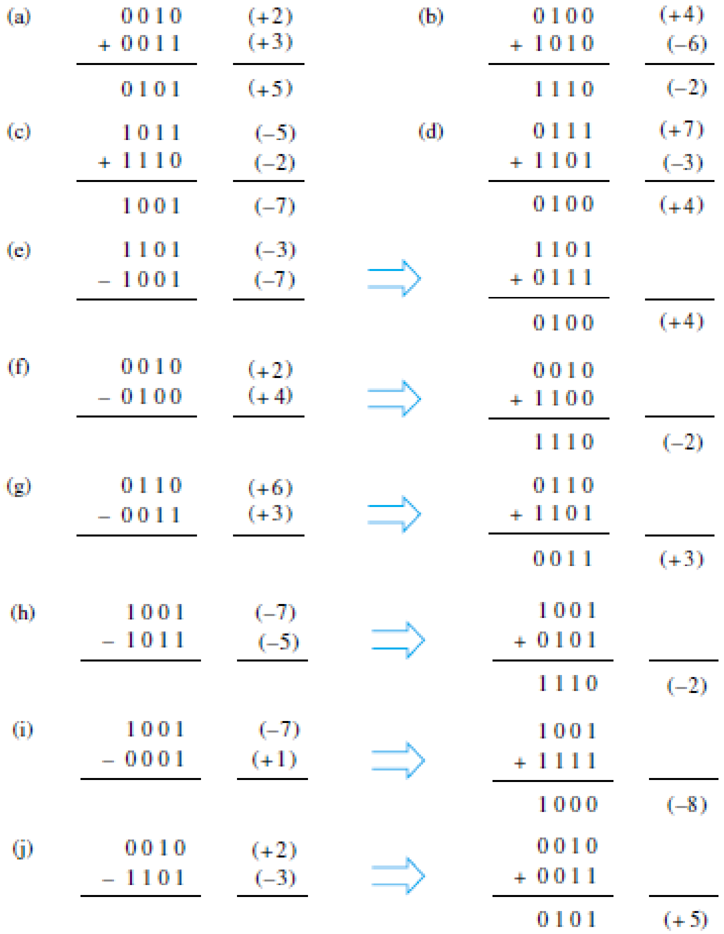

(a) 0010 (+2) + 0011 (+3) = 0101 (+5)
(b) 0100 (+4) + 1010 (−6) = 1110 (−2)
(c) 1011 (−5) + 1110 (−2) = 1001 (−7)
(d) 0111 (+7) + 1101 (−3) = 0100 (+4)
(e) 1101 (−3) − 1001 (−7) ⇒ 1101 + 0111 = 0100 (+4)
(f) 0010 (+2) − 0100 (+4) ⇒ 0010 + 1100 = 1110 (−2)
(g) 0110 (+6) − 0011 (+3) ⇒ 0110 + 1101 = 0011 (+3)
(h) 1001 (−7) − 1011 (−5) ⇒ 1001 + 0101 = 1110 (−2)
(i) 1001 (−7) − 0001 (+1) ⇒ 1001 + 1111 = 1000 (−8)
(j) 0010 (+2) − 1101 (−3) ⇒ 0010 + 0011 = 0101 (+5)

Figure 1.6 2's-complement Add and Subtract operations.

ADDITION & SUBTRACTION OF SIGNED NUMBERS

n-BIT RIPPLE CARRY ADDER

- A cascaded connection of n full-adder blocks can be used to add 2-bit numbers.
- Since carries must propagate (or ripple) through cascade, the configuration is called an n-bit ripple carry adder (Figure 9.1).

x_i	y_i	Carry-in c_i	Sum s_i	Carry-out c_{i+1}
0	0	0	0	0
0	0	1	1	0
0	1	0	1	0
0	1	1	0	1
1	0	0	1	0
1	0	1	0	1
1	1	0	0	1
1	1	1	1	1

$$s_i = \overline{x}_i \overline{y}_i c_i + \overline{x}_i y_i \overline{c}_i + x_i \overline{y}_i \overline{c}_i + x_i y_i c_i = x_i \oplus y_i \oplus c_i$$

$$c_{i+1} = y_i c_i + x_i c_i + x_i y_i$$

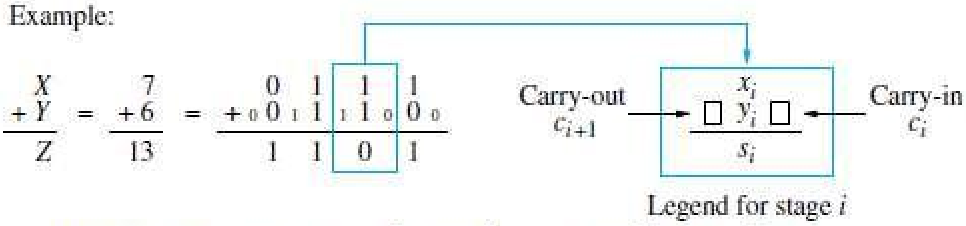
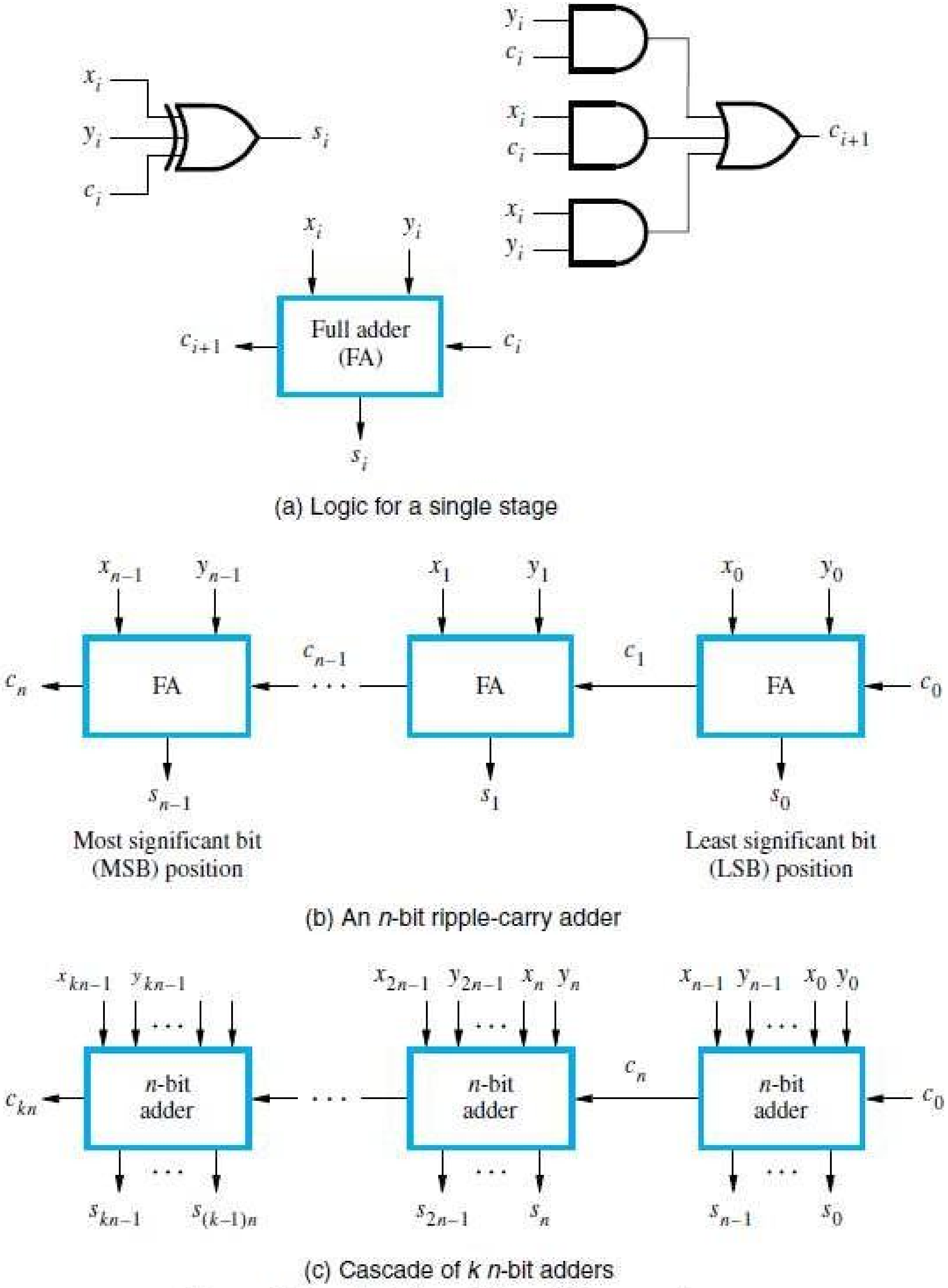


Figure 9.1 Logic specification for a stage of binary addition.


x_i
y_i
c_i
s_i
c_{i+1}
Full adder (FA)
FA
x_{n-1} y_{n-1} x_1 y_1 x_0 y_0
c_n c_{n-1} c_1 c_0
s_{n-1} s_1 s_0
Most significant bit (MSB) position
Least significant bit (LSB) position
n-bit adder
x_{kn-1} y_{kn-1} x_{2n-1} y_{2n-1} x_n y_n x_{n-1} y_{n-1} x_0 y_0
c_{kn} c_n c_0
s_{kn-1} $s_{(k-1)n}$ s_{2n-1} s_n s_{n-1} s_0

(a) Logic for a single stage

(b) An n-bit ripple-carry adder

(c) Cascade of k n-bit adders

Figure 9.2 Logic for addition of binary numbers.

ADDITION/SUBTRACTION LOGIC UNIT

- The n-bit adder can be used to add 2's complement numbers X and Y (Figure 9.3).
- **Overflow** can only occur when the signs of the 2 operands are the same.
- In order to perform the subtraction operation X-Y on 2's complement numbers X and Y; we form the 2's complement of Y and add it to X.
- Addition or subtraction operation is done based on value applied to the Add/Sub input control-line.
- Control-line=0 for addition, applying the Y vector unchanged to one of the adder inputs.
  Control-line=1 for subtraction, the Y vector is 2's complemented.

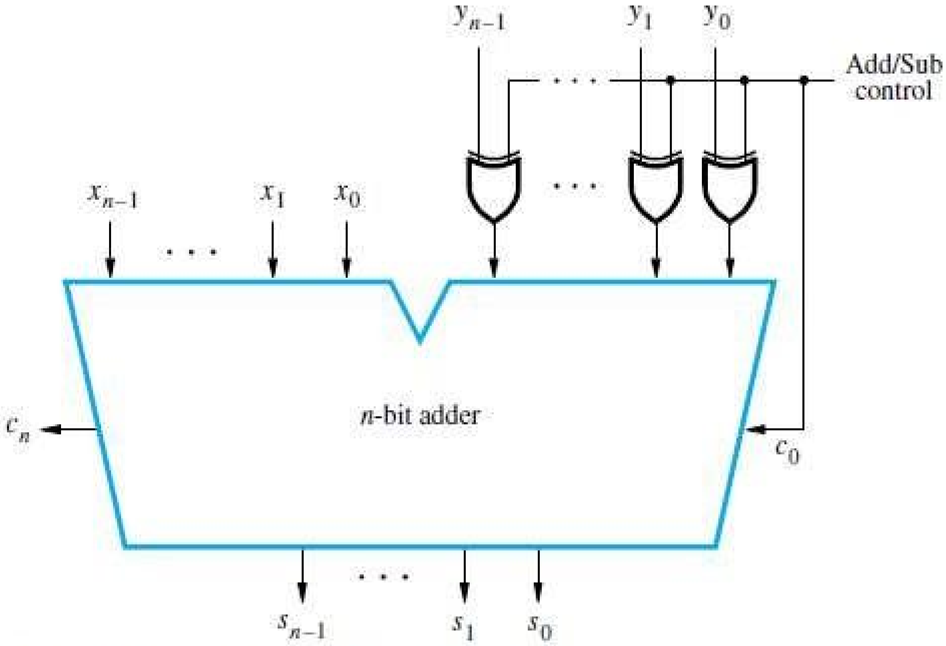


Figure 9.3 Binary addition/subtraction logic circuit.

DESIGN OF FAST ADDERS

- **Drawback of ripple carry adder:** If the adder is used to implement the addition/subtraction, all sum bits are available in 2n gate delays.
- Two approaches can be used to reduce delay in adders:
  1) Use the fastest possible electronic-technology in implementing the ripple-carry design.
  2) Use an augmented logic-gate network structure.

CARRY-LOOKAHEAD ADDITIONS

- The logic expression for s_i(sum) and c_{i+1}(carry-out) of stage i are

  $s_i=x_i+y_i+c_i$ ------(1) $c_{i+1}=x_iy_i+x_ic_i+y_ic_i$ ------(2)

- Factoring (2) into

  $c_{i+1}=x_iy_i+(x_i+y_i)c_i$

  we can write

  $c_{i+1}=G_i+P_iC_i$ where $G_i=x_iy_i$ and $P_i=x_i+y_i$

- The expressions G_i and P_i are called generate and propagate functions (Figure 9.4).
- If $G_i=1$, then $c_{i+1}=1$, independent of the input carry c_i. This occurs when both x_i and y_i are 1. Propagate function means that an input-carry will produce an output-carry when either $x_i=1$ or $y_i=1$.
- All G_i and P_i functions can be formed independently and in parallel in one logic-gate delay.
- Expanding c_i terms of i-1 subscripted variables and substituting into the c_{i+1} expression, we obtain

  $c_{i+1}=G_i+P_iG_{i-1}+P_iP_{i-1}G_{i-2.} \ . \ . \ . \ . \ . \ .+P_1G_0+P_iP_{i-1} \ . \ . \ . \ P_0c_0$

- Conclusion: Delay through the adder is 3 gate delays for all carry-bits &
  4 gate delays for all sum-bits.
- Consider the design of a 4-bit adder. The carries can be implemented as

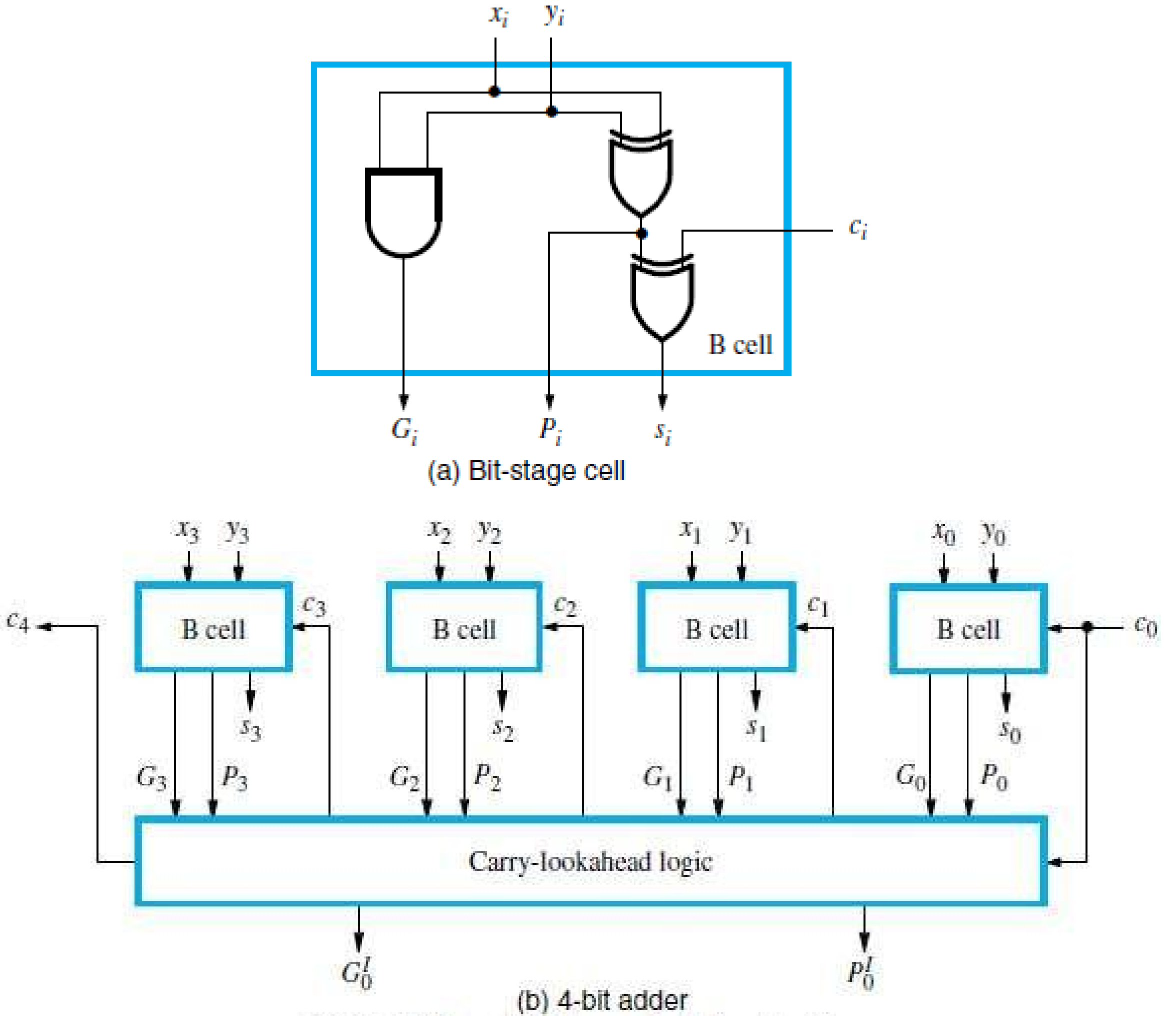
  $c_1=G_0+P_0c_0$

  $c_2=G_1+P_1G_0+P_1P_0c_0$

  $c_3=G_2+P_2G_1+P_2P_1G_0+P_2P_1P_0c_0$

  $c_4=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0+P_3P_2P_1P_0c_0$

- The carries are implemented in the block labeled carry-lookahead logic. An adder implemented in this form is called a **Carry-Lookahead Adder**.
- Limitation: If we try to extend the carry-lookahead adder for longer operands, we run into a problem of gate fan-in constraints.



(a) Bit-stage cell

(b) 4-bit adder

Figure 9.4 A 4-bit carry-lookahead adder.

HIGHER-LEVEL GENERATE & PROPAGATE FUNCTIONS

- 16-bit adder can be built from four 4-bit adder blocks (Figure 9.5).
- These blocks provide new output functions defined as G_k and P_k,
  where k=0 for the first 4-bit block,
  k=1 for the second 4-bit block and so on.
- In the first block,
  $P_0=P_3P_2P_1P_0$
  &
  $G_0=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0$
- The first-level G_i and P_i functions determine whether bit stage i generates or propagates a carry, and the second level G_k and P_k functions determine whether block k generates or propagates a carry.
- Carry c_{16} is formed by one of the carry-lookahead circuits as
  $c_{16}=G_3+P_3G_2+P_3P_2G_1+P_3P_2P_1G_0+P_3P_2P_1P_0c_0$
- Conclusion: All carries are available 5 gate delays after X, Y and c_0 are applied as inputs.

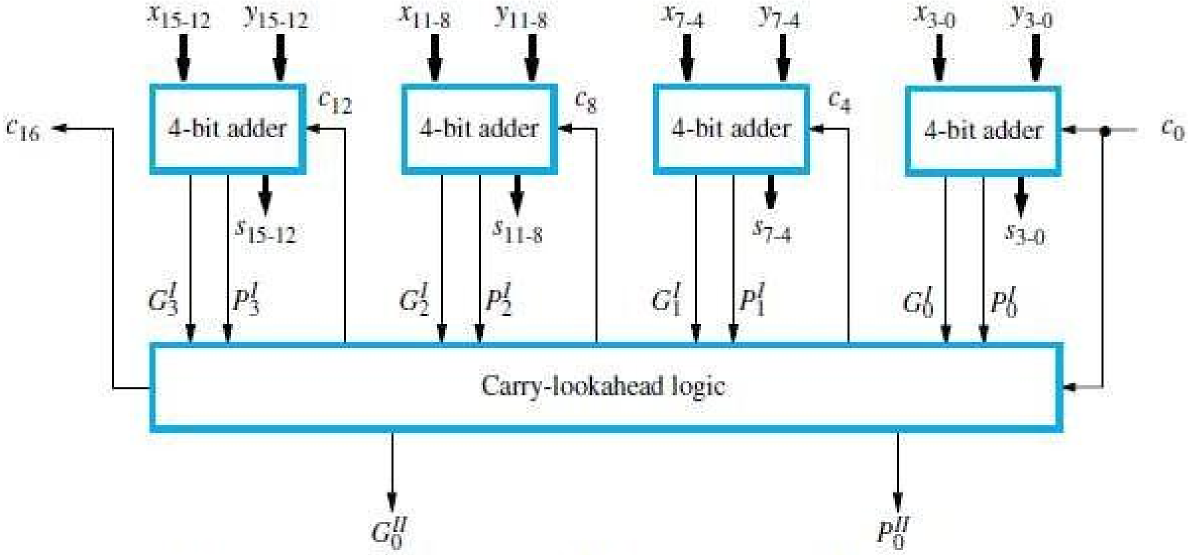
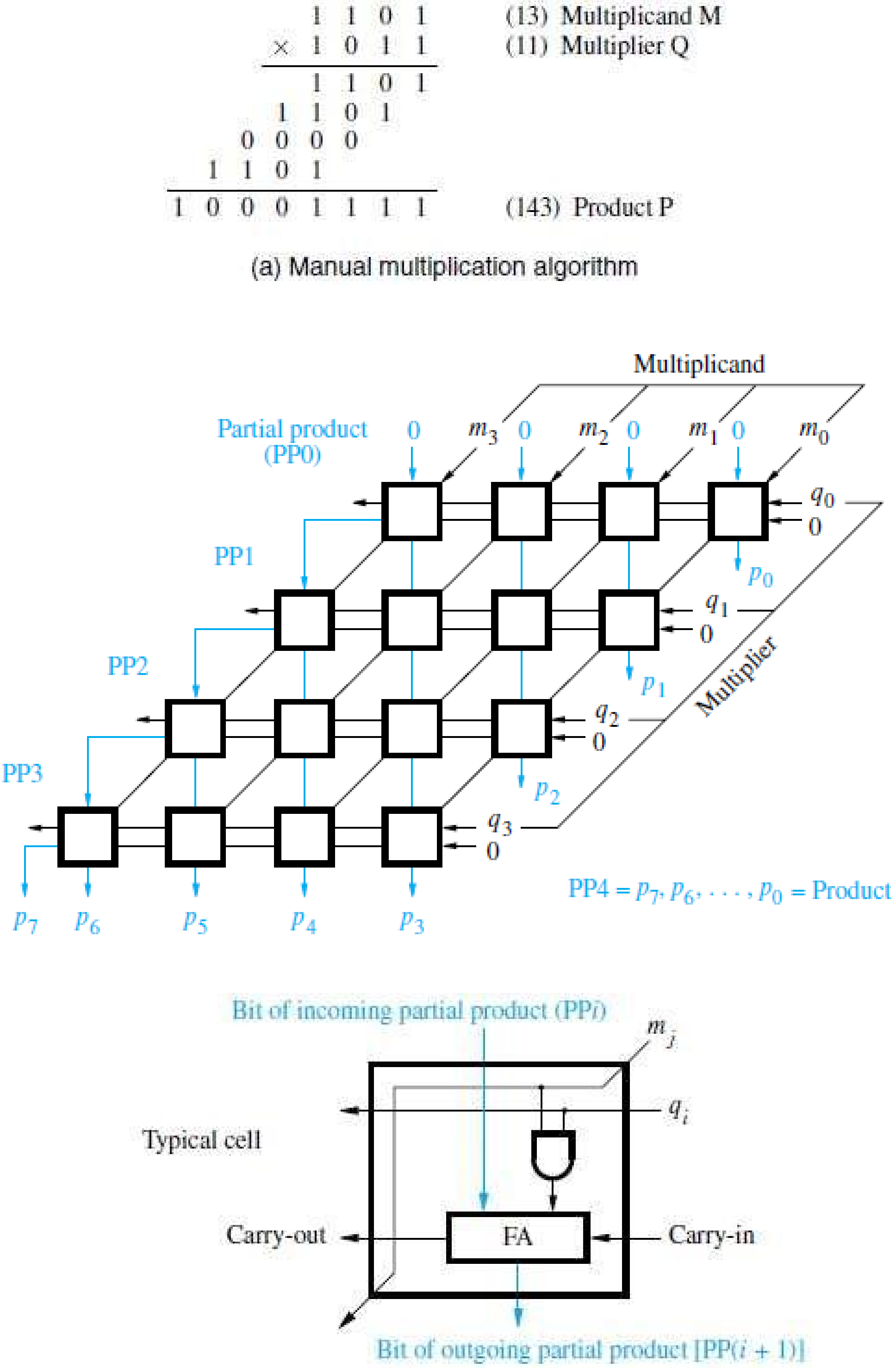


Figure 9.5 A 16-bit carry-lookahead adder built from 4-bit adders (see Figure 9.4*b*).

MULTIPLICATION OF POSITIVE NUMBERS


1 1 0 1 (13) Multiplicand M
× 1 0 1 1 (11) Multiplier Q
1 1 0 1
1 1 0 1
0 0 0 0
1 1 0 1
1 0 0 0 1 1 1 1 (143) Product P
Multiplicand
Partial product (PP0)
0
m_3
m_2
m_1
m_0
q_0
q_1
q_2
q_3
PP1
PP2
PP3
p_0
p_1
p_2
p_3
p_4
p_5
p_6
p_7
Multiplier
PP4 = $p_7, p_6, \ldots, p_0$ = Product
Bit of incoming partial product (PPi)
m_j
q_i
Typical cell
Carry-out
FA
Carry-in
Bit of outgoing partial product [PP(i + 1)]

(a) Manual multiplication algorithm

(b) Array implementation

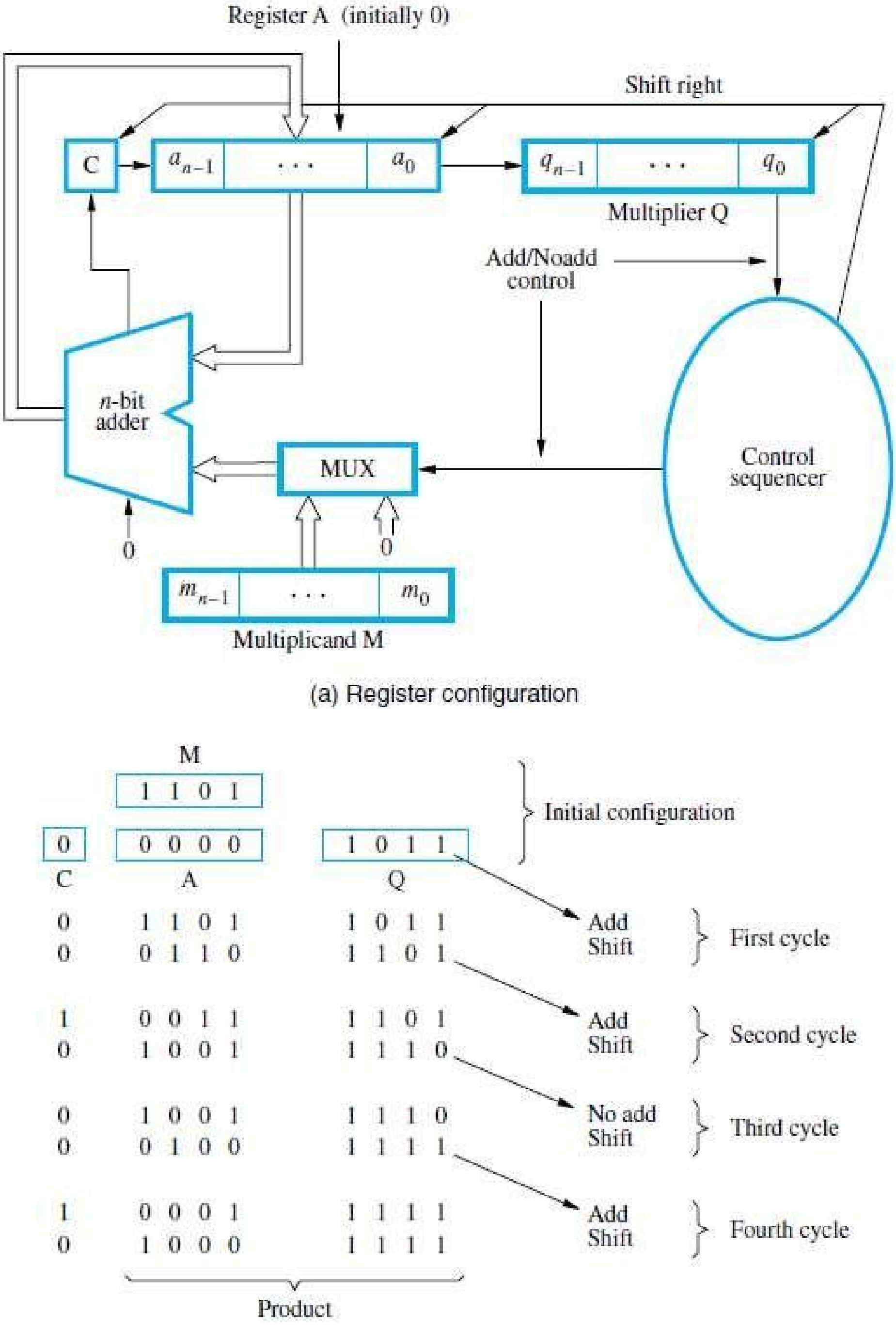
Figure 9.6 Array multiplication of unsigned binary operands.

ARRAY MULTIPLICATION

- The main component in each cell is a full adder(FA)..
- The AND gate in each cell determines whether a multiplicand bit m_j, is added to the incoming partial-product bit, based on the value of the multiplier bit q_i (Figure 9.6).

SEQUENTIAL CIRCUIT BINARY MULTIPLIER

- Registers A and Q combined hold PP_i(partial product)
  while the multiplier bit q_i generates the signal Add/Noadd.
- The carry-out from the adder is stored in flip-flop C (Figure 9.7).
- Procedure for multiplication:
  1) Multiplier is loaded into register Q,
     Multiplicand is loaded into register M and
     C & A are cleared to 0.
  2) If q_0=1, add M to A and store sum in A. Then C, A and Q are shifted right one bit-position.
     If q_0=0, no addition performed and C, A & Q are shifted right one bit-position.
  3) After n cycles, the high-order half of the product is held in register A and
     the low-order half is held in register Q.



(a) Register configuration

C	A	Q		
	M: 1 1 0 1			Initial configuration
0	0 0 0 0	1 0 1 1		
0	1 1 0 1	1 0 1 1	Add	First cycle
0	0 1 1 0	1 1 0 1	Shift	
1	0 0 1 1	1 1 0 1	Add	Second cycle
0	1 0 0 1	1 1 1 0	Shift	
0	1 0 0 1	1 1 1 0	No add	Third cycle
0	0 1 0 0	1 1 1 1	Shift	
1	0 0 0 1	1 1 1 1	Add	Fourth cycle
0	1 0 0 0	1 1 1 1	Shift	

(b) Multiplication example

Figure 9.7 Sequential circuit binary multiplier.

SIGNED OPERAND MULTIPLICATION

BOOTH ALGORITHM

- This algorithm
  - → generates a 2n-bit product
  - → treats both positive & negative 2's-complement n-bit operands uniformly(Figure 9.9-9.12).
- Attractive feature: This algorithm achieves some efficiency in the number of addition required when the multiplier has a few large blocks of 1s.
- This algorithm suggests that we can reduce the number of operations required for multiplication by representing multiplier as a difference between 2 numbers.

  For e.g. multiplier(Q) 14(001110) can be represented as

  ```
   010000 (16)
  -000010 (2)
  001110 (14)
  ```
- Therefore, product P=M*Q can be computed by adding 2^4 times the M to the 2's complement of 2^1 times the M.

```
                       0  1  0  1  1  0  1
                       0  0 +1 +1 +1 +1  0
                     ---------------------
                       0  0  0  0  0  0  0
                    0  1  0  1  1  0  1
                 0  1  0  1  1  0  1
              0  1  0  1  1  0  1
           0  1  0  1  1  0  1
        0  0  0  0  0  0  0
     0  0  0  0  0  0  0
  ------------------------------------------
  0  0  0  1  0  1  0  1  0  0  0  1  1  0
```

```
                       0  1  0  1  1  0  1
                       0 +1  0  0  0 -1  0
                     ---------------------
  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  1  1  1  1  1  1  1  0  1  0  0  1  1      ←—— 2's complement of
  0  0  0  0  0  0  0  0  0  0  0  0              the multiplicand
  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0
  0  0  0  1  0  1  1  0  1
  0  0  0  0  0  0  0  0
  ------------------------------------------
  0  0  0  1  0  1  0  1  0  0  0  1  1  0
```

Figure 9.9 Normal and Booth multiplication schemes.

```
 0  0  1  0  1  1  0  0  1  1  1  0  1  0  1  1  0  0
                             ⇓
 0 +1 -1 +1  0 -1  0 +1  0  0 -1 +1 -1 +1  0 -1  0  0
```

Figure 9.10 Booth recoding of a multiplier.

```
    0  1  1  0  1   (+13)          0  1  1  0  1
 ×  1  1  0  1  0   (−6)    ⇒      0 -1 +1 -1  0
 ----------------               ---------------
                   0  0  0  0  0  0  0  0  0  0
                   1  1  1  1  1  0  0  1  1
                   0  0  0  0  1  1  0  1
                   1  1  1  0  0  1  1
                   0  0  0  0  0  0
                ------------------------------
                   1  1  1  0  1  1  0  0  1  0   (−78)
```

Figure 9.11 Booth multiplication with a negative multiplier.

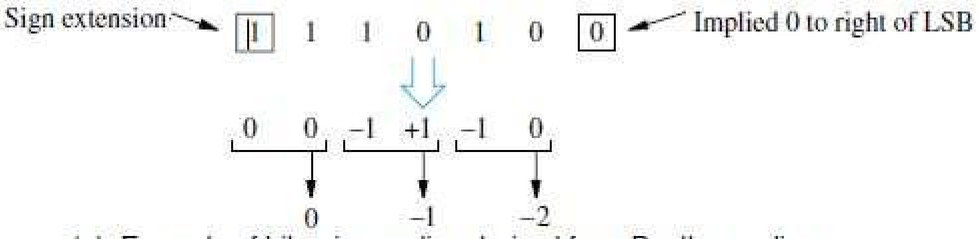
Multiplier		Version of multiplicand selected by bit *i*
Bit *i*	Bit *i* − 1	
0	0	0 × M
0	1	+ 1 × M
1	0	− 1 × M
1	1	0 × M

Figure 9.12 Booth multiplier recoding table.

FAST MULTIPLICATION

BIT-PAIR RECODING OF MULTIPLIERS

- This method
  - → derived from the booth algorithm
  - → reduces the number of summands by a factor of 2
- Group the Booth-recoded multiplier bits in pairs. (Figure 9.14 & 9.15).
- The pair (+1 -1) is equivalent to the pair (0 +1).



(a) Example of bit-pair recoding derived from Booth recoding

Multiplier bit-pair		Multiplier bit on the right	Multiplicand selected at position i
$i+1$	i	$i-1$	
0	0	0	0 × M
0	0	1	+ 1 × M
0	1	0	+ 1 × M
0	1	1	+ 2 × M
1	0	0	− 2 × M
1	0	1	− 1 × M
1	1	0	− 1 × M
1	1	1	0 × M

(b) Table of multiplicand selection decisions

Figure 9.14 Multiplier bit-pair recoding.

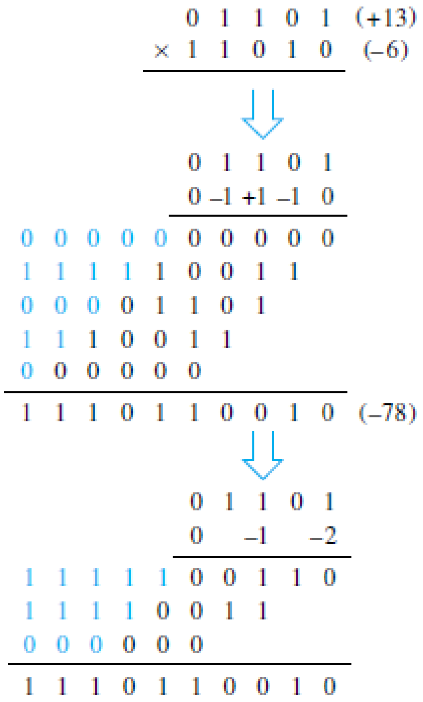
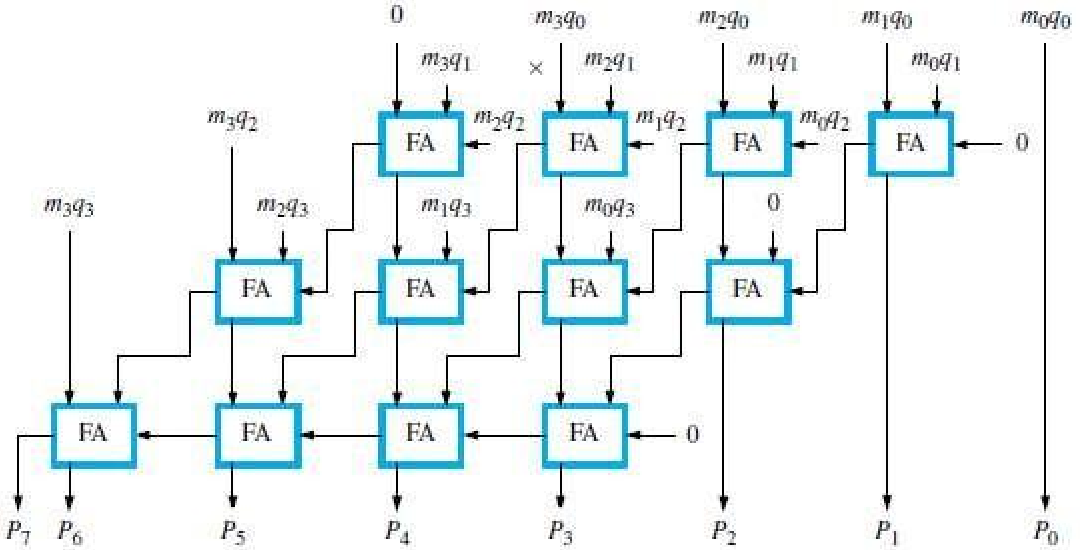


Figure 9.15 Multiplication requiring only $n/2$ summands.

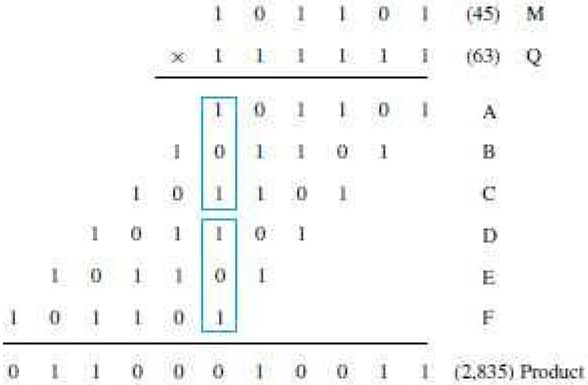
CARRY-SAVE ADDITION OF SUMMANDS

- Consider the array for 4*4 multiplication. (Figure 9.16 & 9.18).
- Instead of letting the carries ripple along the rows, they can be "saved" and introduced into the next row, at the correct weighted positions.



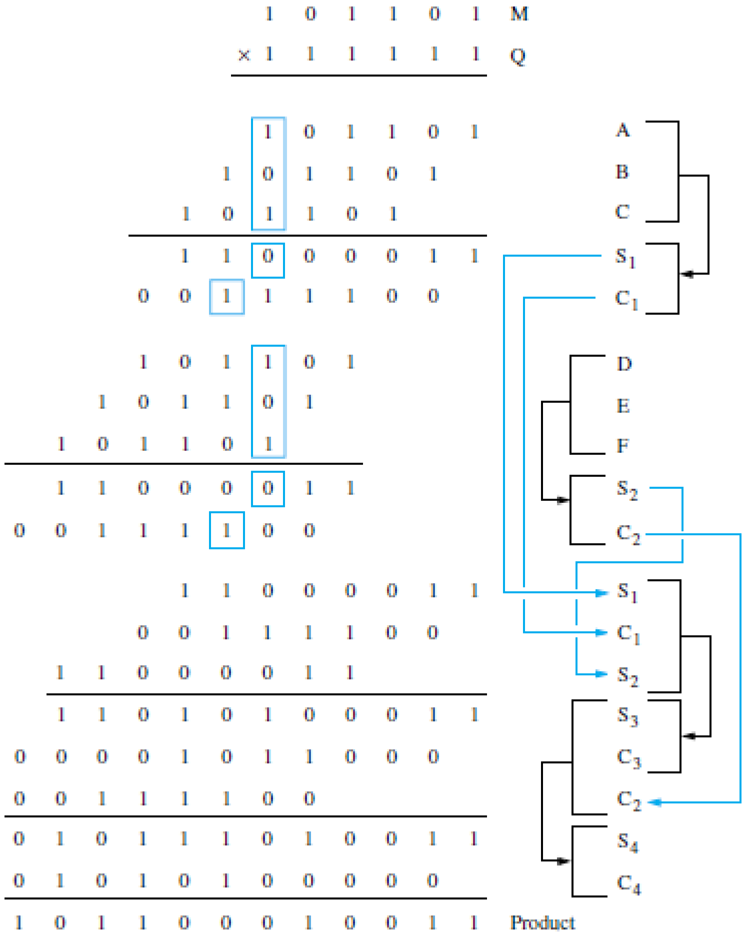
(b) Carry-save array

Figure 9.16 carry-save arrays for a 4 × 4 multiplier.



```
                    1   0   1   1   0   1    (45)  M
                ×   1   1   1   1   1   1    (63)  Q
                    1   0   1   1   0   1          A
                1   0   1   1   0   1              B
            1   0   1   1   0   1                  C
        1   0   1   1   0   1                      D
    1   0   1   1   0   1                          E
1   0   1   1   0   1                              F
0   1   1   0   0   0   1   0   0   1   1    (2,835) Product
```

Figure 9.17 A multiplication example used to illustrate carry-save addition as shown in Figure 9.18.



```
                        1   0   1   1   0   1    M
                    ×   1   1   1   1   1   1    Q

                        1   0   1   1   0   1    A
                    1   0   1   1   0   1        B
                1   0   1   1   0   1            C
                1   1   0   0   0   0   1   1    S1
            0   0   1   1   1   1   0   0        C1

            1   0   1   1   0   1                D
        1   0   1   1   0   1                    E
    1   0   1   1   0   1                        F
    1   1   0   0   0   0   1   1                S2
0   0   1   1   1   1   0   0                    C2

                1   1   0   0   0   0   1   1    S1
            0   0   1   1   1   1   0   0        C1
    1   1   0   0   0   0   1   1                S2
    1   1   0   1   0   1   0   0   0   1   1    S3
0   0   0   0   1   0   1   1   0   0   0        C3
0   0   1   1   1   1   0   0                    C2
0   1   0   1   1   1   0   1   0   0   1   1    S4
0   1   0   1   0   1   0   0   0   0   0        C4
1   0   1   1   0   0   0   1   0   0   1   1    Product
```

Figure 9.18 The multiplication example from Figure 9.17 performed using carry-save addition.

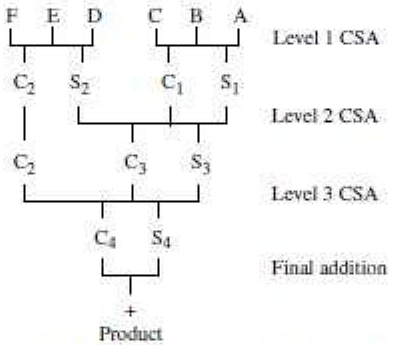


Figure 9.19 Schematic representation of the carry-save addition operations in Figure 9.18.

- The full adder is input with three partial bit products in the first row.
- Multiplication requires the addition of several summands.
- CSA speeds up the addition process.
- Consider the array for 4x4 multiplication shown in fig 9.16.
- First row consisting of just the AND gates that implement the bit products m_3q_0, m_2q_0, m_1q_0 and m_0q_0.
- The delay through the carry-save array is somewhat less than delay through the ripple-carry array. This is because the S and C vector outputs from each row are produced in parallel in one full-adder delay.
- Consider the addition of many summands in fig 9.18.
- Group the summands in threes and perform carry-save addition on each of these groups in parallel to generate a set of S and C vectors in one full-adder delay
- Group all of the S and C vectors into threes, and perform carry-save addition on them, generating a further set of S and C vectors in one more full-adder delay
- Continue with this process until there are only two vectors remaining
- They can be added in a RCA or CLA to produce the desired product.
- When the number of summands is large, the time saved is proportionally much greater.
- Delay: AND gate + 2 gate/CSA level + CLA gate delay, Eg., 6 bit number require 15 gate delay, array 6x6 require 6(n-1)-1 = 29 gate Delay.
- In general, CSA takes 1.7 log_2k-1.7 levels of CSA to reduce k summands.

INTEGER DIVISION

- An n-bit positive-divisor is loaded into register M.
  An n-bit positive-dividend is loaded into register Q at the start of the operation.
  Register A is set to 0 (Figure 9.21).
- After division operation, the n-bit quotient is in register Q, and
  the remainder is in register A.

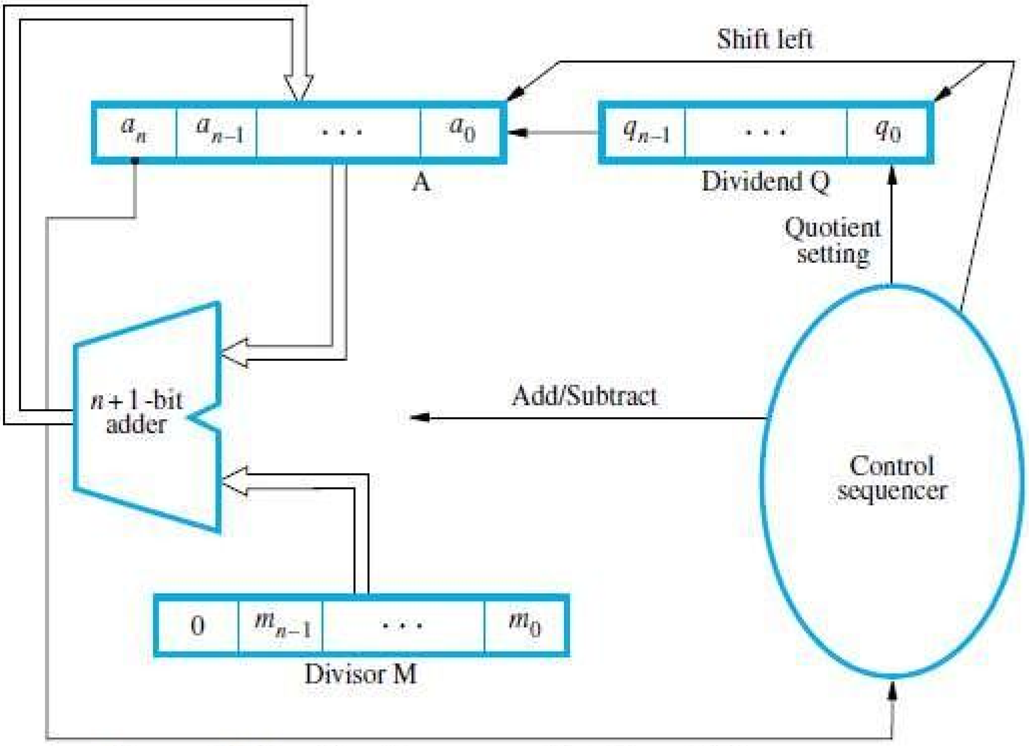


Figure 9.23 Circuit arrangement for binary division.

```
       21                    10101
  13 ) 274        1101 ) 100010010
       26                1101
        14                10000
        13                 1101
         1                  1110
                            1101
                               1
```

Figure 9.22 Longhand division examples.

NON-RESTORING DIVISION

- Procedure:

  Step 1: Do the following n times

  i) If the sign of A is 0, shift A and Q left one bit position and subtract M from A; otherwise, shift A and Q left and add M to A (Figure 9.23).

  ii) Now, if the sign of A is 0, set q_0 to 1; otherwise set q_0 to 0.

  Step 2: If the sign of A is 1, add M to A (restore).

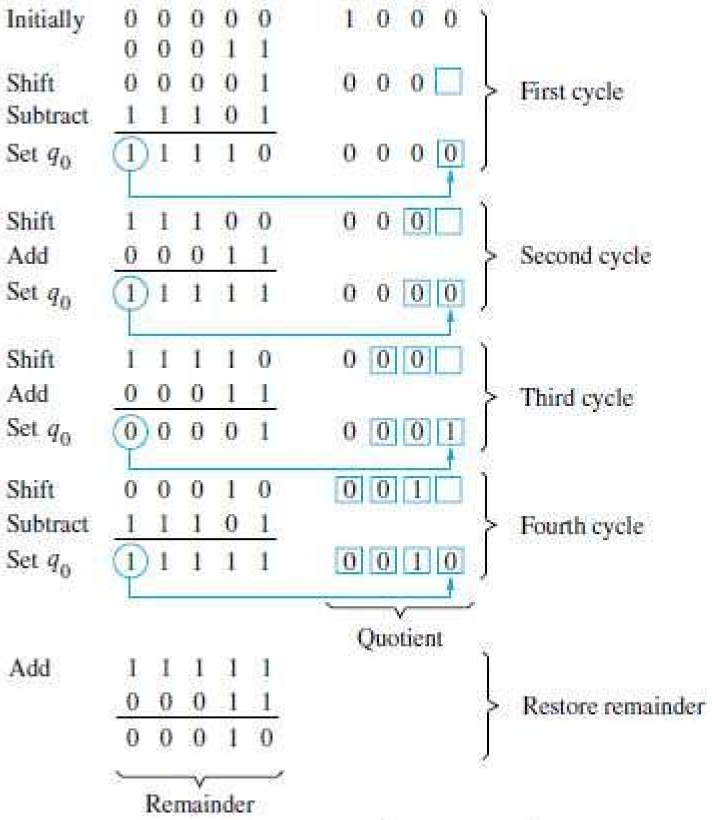


Figure 9.25 A non-restoring division example.

RESTORING DIVISION

- Procedure: Do the following n times

  1) Shift A and Q left one binary position (Figure 9.22).
  2) Subtract M from A, and place the answer back in A
  3) If the sign of A is 1, set q_0 to 0 and add M back to A(restore A).
     If the sign of A is 0, set q_0 to 1 and no restoring done.

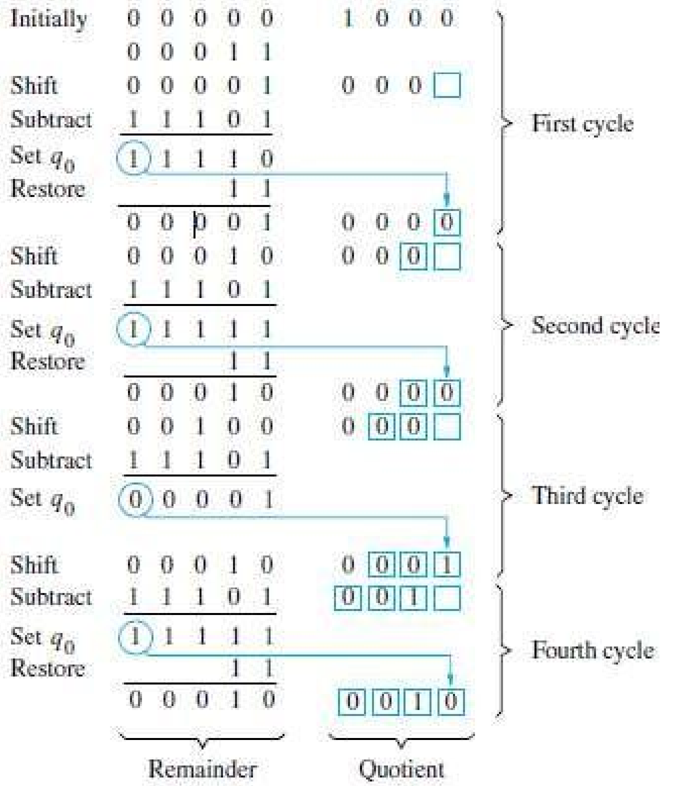


Figure 9.24 A restoring division example.

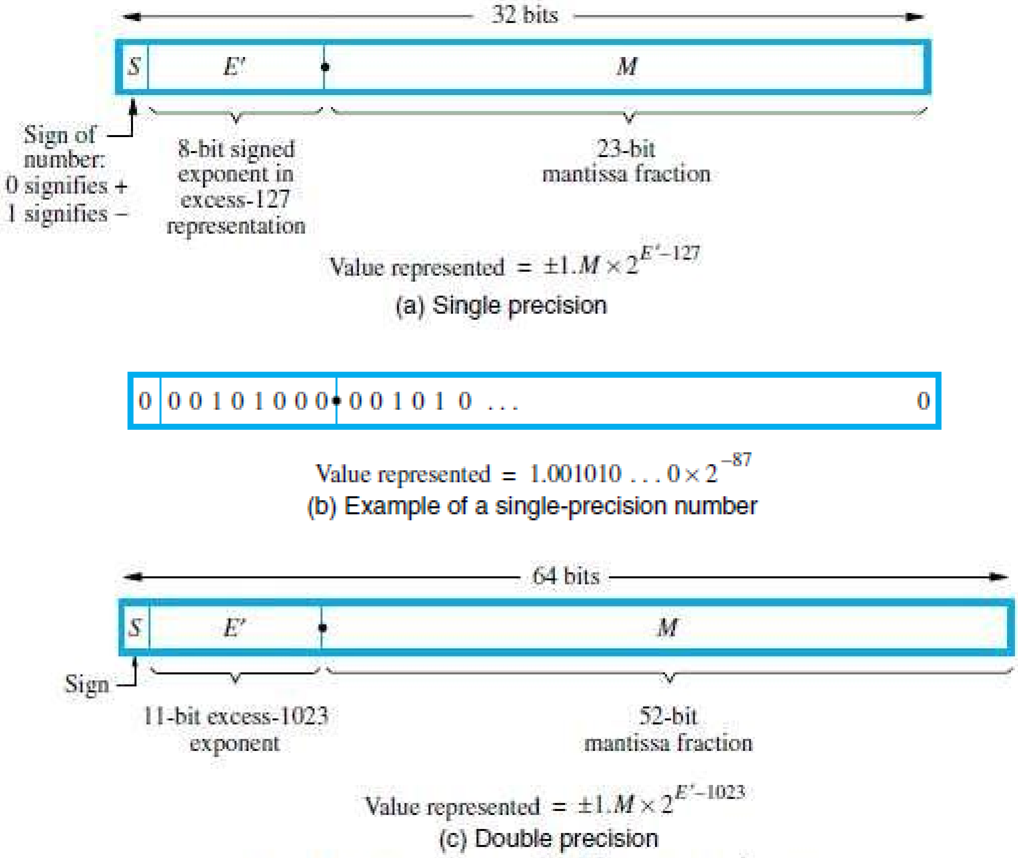
FLOATING-POINT NUMBERS & OPERATIONS

IEEE STANDARD FOR FLOATING POINT NUMBERS

- Single precision representation occupies a single 32-bit word.

  The scale factor has a range of 2^{-126} to 2^{+127}(which is approximately equal to 10^{+38}).
- The 32 bit word is divided into 3 fields: sign(1 bit), exponent(8 bits) and mantissa(23 bits).
- Signed exponent=E.

  Unsigned exponent E'=E+127. Thus, E' is in the range 0<E'<255.
- The last 23 bits represent the mantissa. Since binary normalization is used, the MSB of the mantissa is always equal to 1. (M represents fractional-part).
- The 24-bit mantissa provides a precision equivalent to about 7 decimal-digits (Figure 9.24).
- Double precision representation occupies a single 64-bit word. And E' is in the range 1<E'<2046.
- The 53-bit mantissa provides a precision equivalent to about 16 decimal-digits.


32 bits
S
E'
M
Sign of number: 0 signifies + 1 signifies −
8-bit signed exponent in excess-127 representation
23-bit mantissa fraction
0
0 0 1 0 1 0 0 0
0 0 1 0 1 0 ...
0
64 bits
S
E'
M
Sign
11-bit excess-1023 exponent
52-bit mantissa fraction

Value represented $= \pm 1.M \times 2^{E'-127}$

(a) Single precision

Value represented $= 1.001010\ldots0 \times 2^{-87}$

(b) Example of a single-precision number

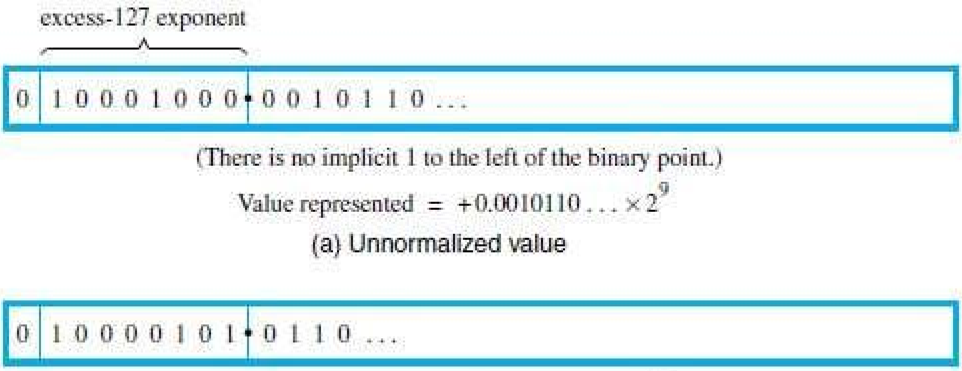
Value represented $= \pm 1.M \times 2^{E'-1023}$

(c) Double precision

Figure 9.26 IEEE standard floating-point formats.

NORMALIZATION

- When the decimal point is placed to the right of the first(non zero) significant digit, the number is said to be normalized.
- If a number is not normalized, it can always be put in normalized form by shifting the fraction and adjusting the exponent. As computations proceed, a number that does not fall in the representable range of normal numbers might be generated.
- In single precision, it requires an exponent less than -126 (underflow) or greater than +127 (overflow). Both are exceptions that need to be considered.



(There is no implicit 1 to the left of the binary point.)

Value represented = $+0.0010110\ldots \times 2^{9}$

(a) Unnormalized value

Value represented = $+1.0110\ldots \times 2^{6}$

(b) Normalized version

Figure 9.27 Floating-point normalization in IEEE single-precision format.

SPECIAL VALUES

- The end values 0 and 255 of the excess-127 exponent E' are used to represent special values.
- When E'=0 and the mantissa fraction m is zero, the value exact 0 is represented.
- When E'=255 and M=0, the value ∞ is represented, where ∞ is the result of dividing a normal number by zero.
- when E'=0 and M!=-, denormal numbers are represented. Their value is $\pm 0.M X2^{-126}$
- When E'=255 and M!=0, the value represented is called not a number(NaN). A NaN is the result of performing an invalied operation such as 0/0 or $\sqrt{0}$.

ARITHMETIC OPERATIONS ON FLOATING-POINT NUMBERS

Multiply Rule

1) Add the exponents & subtract 127.
2) Multiply the mantissas & determine sign of the result.
3) Normalize the resulting value if necessary.

Divide Rule

1) Subtract the exponents & add 127.
2) Divide the mantissas & determine sign of the result.
3) Normalize the resulting value if necessary.

Add/Subtract Rule

1) Choose the number with the smaller exponent & shift its mantissa right a number of steps equal to the difference in exponents(n).
2) Set exponent of the result equal to larger exponent.
3) Perform addition/subtraction on the mantissas & determine sign of the result.
4) Normalize the resulting value if necessary.

IMPLEMENTING FLOATING-POINT OPERATIONS

- First compare exponents to determine how far to shift the mantissa of the number with the smaller exponent.
- The shift-count value n
  → is determined by 8 bit subtractor &
  → is sent to SHIFTER unit.
- In step 1, sign is sent to SWAP network (Figure 9.26).
  If sign=0, then $E_A > E_B$ and mantissas M_A & M_B are sent straight through SWAP network.
  If sign=1, then $E_A < E_B$ and the mantissas are swapped before they are sent to SHIFTER.
- In step 2, 2:! MUX is used. The exponent of result E is tentatively determined as E_A if $E_A > E_B$ or
  E_B if $E_A < E_B$
- In step 3, CONTROL logic
  → determines whether mantissas are to be added or subtracted.
  → determines sign of the result.
- In step 4, result of step 3 is normalized. The number of leading zeros in M determines number of bit shifts(X) to be applied to M.

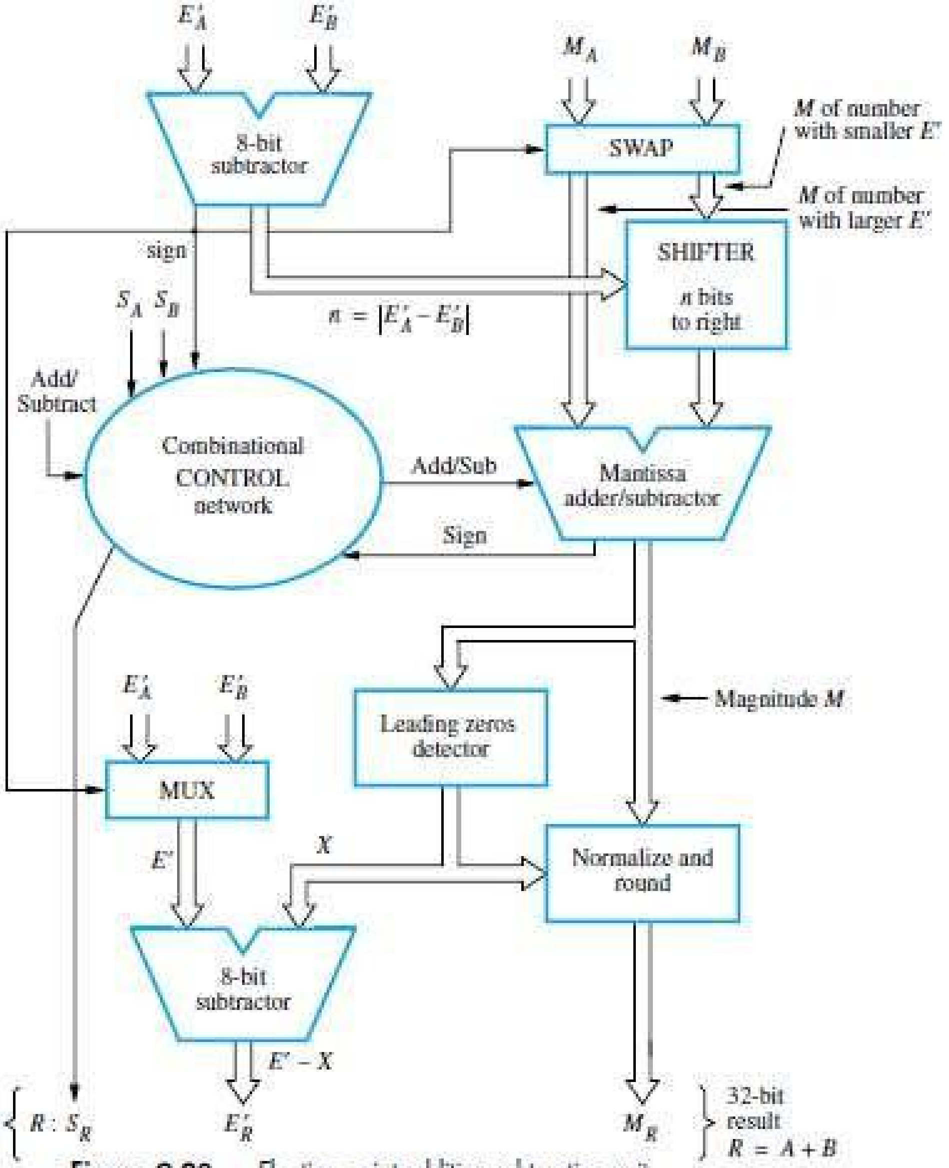


Figure 9.28 Floating-point addition-subtraction unit.

Problem 1:

Represent the decimal values 5, -2, 14, -10, 26, -19, 51 and -43 as signed 7-bit numbers in the following binary formats:

(a) sign-and-magnitude
(b) 1's-complement
(c) 2's-complement

Solution:

The three binary representations are given as:

Decimal values	Sign-and-magnitude representation	1's-complement representation	2's-complement representation
5	0000101	0000101	0000101
−2	1000010	1111101	1111110
14	0001110	0001110	0001110
−10	1001010	1110101	1110110
26	0011010	0011010	0011010
−19	1010011	1101100	1101101
51	0110011	0110011	0110011
−43	1101011	1010100	1010101

Problem 2:

(a) Convert the following pairs of decimal numbers to 5-bit 2's-complement numbers, then add them. State whether or not overflow occurs in each case.

a) 5 and 10 b) 7 and 13
c) −14 and 11 d) −5 and 7
e) −3 and −8

(b) Repeat Problem 1.7 for the subtract operation, where the second number of each pair is to be subtracted from the first number. State whether or not overflow occurs in each case.

Solution:

(a)

$$\begin{array}{lr@{\qquad}lr@{\qquad}lr}
(a) & 00101 & (b) & 00111 & (c) & 10010 \\
 & +\,01010 & & +\,01101 & & +\,01011 \\ \cline{2-2}\cline{4-4}\cline{6-6}
 & 01111 & & 10100 & & 11101 \\
 & \text{no overflow} & & \text{overflow} & & \text{no overflow}
\end{array}$$

$$\begin{array}{lr@{\qquad}lr@{\qquad}lr}
(d) & 11011 & (e) & 11101 & (f) & 10110 \\
 & +\,00111 & & +\,11000 & & +\,10011 \\ \cline{2-2}\cline{4-4}\cline{6-6}
 & 00010 & & 10101 & & 01001 \\
 & \text{no overflow} & & \text{no overflow} & & \text{overflow}
\end{array}$$

(b) To subtract the second number, form its 2's-complement and add it to the first number.

$$\begin{array}{lr@{\qquad}lr@{\qquad}lr}
(a) & 00101 & (b) & 00111 & (c) & 10010 \\
 & +\,10110 & & +\,10011 & & +\,10101 \\ \cline{2-2}\cline{4-4}\cline{6-6}
 & 11011 & & 11010 & & 00111 \\
 & \text{no overflow} & & \text{no overflow} & & \text{overflow}
\end{array}$$

$$\begin{array}{lr@{\qquad}lr@{\qquad}lr}
(d) & 11011 & (e) & 11101 & (f) & 10110 \\
 & +\,11001 & & +\,01000 & & +\,01101 \\ \cline{2-2}\cline{4-4}\cline{6-6}
 & 10100 & & 00101 & & 00011 \\
 & \text{no overflow} & & \text{no overflow} & & \text{no overflow}
\end{array}$$

Problem 3:
Perform following operations on the 6-bit signed numbers using 2's complement representation system. Also indicate whether overflow has occurred.

$$\begin{array}{rrr}
010110 & 101011 & 111111 \\
\underline{+001001} & \underline{+100101} & \underline{+000111} \\
011001 & 110111 & 010101 \\
\underline{+010000} & \underline{+111001} & \underline{+101011} \\
010110 & 111110 & 100001 \\
\underline{-011111} & \underline{-100101} & \underline{-011101} \\
111111 & 000111 & 011010 \\
\underline{-000111} & \underline{-111000} & \underline{-100010}
\end{array}$$

Solution:

$$\begin{array}{rrrrrr}
010110 & (+22) & 101011 & (-21) & 111111 & (-1) \\
\underline{+001001} & \underline{+(+9)} & \underline{+100101} & \underline{+(-27)} & \underline{+000111} & \underline{+(+7)} \\
011111 & (+31) & 010000 & (-48) & 000110 & (+6) \\
 & & \text{overflow} & & &
\end{array}$$

$$\begin{array}{rrrrrr}
011001 & (+25) & 110111 & (-9) & 010101 & (+21) \\
\underline{+010000} & \underline{+(+16)} & \underline{+111001} & \underline{+(-7)} & \underline{+101011} & \underline{+(-21)} \\
101001 & (+41) & 110000 & (-16) & 000000 & (0) \\
\text{overflow} & & & & &
\end{array}$$

$$\begin{array}{rrr}
010110 & (+22) & 010110 \\
\underline{-011111} & \underline{-(+31)} & \underline{+100001} \\
 & (-9) & 110111
\end{array}$$

$$\begin{array}{rrr}
111110 & (-2) & 111110 \\
\underline{-100101} & \underline{-(-27)} & \underline{+011011} \\
 & (+25) & 011001
\end{array}$$

$$\begin{array}{rrr}
100001 & (-31) & 100001 \\
\underline{-011101} & \underline{-(+29)} & \underline{+100011} \\
 & (-60) & 000100 \\
 & & \text{overflow}
\end{array}$$

$$\begin{array}{rrr}
111111 & (-1) & 111111 \\
\underline{-000111} & \underline{-(+7)} & \underline{+111001} \\
 & (-8) & 111000
\end{array}$$

$$\begin{array}{rrr}
000111 & (+7) & 000111 \\
\underline{-111000} & \underline{-(-8)} & \underline{+001000} \\
 & (+15) & 001111
\end{array}$$

$$\begin{array}{rrr}
011010 & (+26) & 011010 \\
\underline{-100010} & \underline{-(-30)} & \underline{+011110} \\
 & (+56) & 111000 \\
 & & \text{overflow}
\end{array}$$

Problem 4:

Perform signed multiplication of following 2's complement numbers using Booth's algorithm.

(a) A=010111 and B=110110 (b) A=110011 and B=101100

(c) A=110101 and B=011011 (d) A=001111 and B=001111

(e) A=10100 and B=10101 (f) A=01110 and B=11000

Solution:

```
  010111       +23                               0  1  0  1  1  1
 ×110110      × -10                          ×   0 -1 +1  0 -1  0
 -------      -----                         ---------------------
               -230                                             0
                              1  1  1  1  1  1  0  1  0  0  1
                 sign                                      0
              extension       0  0  0  0  1  0  1  1  1
                              1_1 1_1 1_1 0_2 1_1 0_1 0_1 1
                              -----------------------------------
                              1  1  1  1  0  0  0  1  1  0  1  0
```

```
  110011       -13                               1  1  0  0  1  1
 ×101100      × -20                          ×  -1 +1  0 -1  0  0
 -------      -----                         ---------------------
                260                                             0
                                                             0
                              0  0  0  0  0  0  1  1  0  1
                 sign                                0
              extension       1  1  1  1  0  0  1  1
                              0_1 0_1 0_1 1_1 1_1 0_2 1_1
                              -----------------------------------
                              0  0  0  1  0  0  0  0  0  1  0  0
```

```
  110101       -11                               1  1  0  1  0  1
 ×011011      ×  27                          ×  +1  0 -1 +1  0 -1
 -------      -----                         ---------------------
               -297           0  0  0  0  0  0  0  0  1  0  1  1
                                                             0
                 sign         1  1  1  1  1  1  0  1  0  1
              extension       0  0  0  0  0  1  0  1  1
                                                   0
                              1_1 1_1 1_1 0_1 1_1 0_1 1_1   _1
                              -----------------------------------
                              1  1  1  0  1  1  0  1  0  1  1  1
```

```
  001111        15                               0  0  1  1  1  1
 ×001111      × 15                           ×   0 +1  0  0  0 -1
 -------      -----                         ---------------------
                225           1  1  1  1  1  1  1  1  0  0  0  1
                              0  0  0  0  1  1  1  1
                              -----------------------------------
                              0  0  0  0  1  1  1  0  0  0  0  1
```

```
   1 0 1 0 0 (-12)                        1  0  1  0  0
 X 1 0 1 0 1 (-11)     →                 -1  1 -1  1 -1   (recoded multiplier)
 -----------------               -----------------------
                                 0  0  0  0  0  1  1  0  0
                                 1  1  1  1  0  1  0  0
                                 0  0  0  1  1  0  0
                                 1  1  0  1  0  0
                                 0  1  1  0  0
                                 -------------------------
                                 0  1  0  0  0  0  1  0  0   (+132)
```

```
   0 1 1 1 0 (+14)                     0  1  1  1  0
 X 1 1 0 0 0 (-8)      →               0 -1  0  0  0   (recoded multiplier)
 -----------------          -----------------------------
                            0  0  0  0  0  0  0  0  0  0
                            0  0  0  0  0  0  0  0  0
                            0  0  0  0  0  0  0  0
                            1  1  1  0  0  1  0
                            0  0  0  0  0  0
                            -----------------------------
                            1  1  1  0  0  1  0  0  0  0   (-112)
```

Problem 5:

Perform signed multiplication of following 2’s complement numbers using bit-pair recoding method.

(a) A=010111 and B=110110 (b) A=110011 and B=101100

(c) A=110101 and B=011011 (d) A=001111 and B=001111

Solution:

$$\begin{array}{r} 010111 \\ \times\,110110 \\ \hline \end{array}$$

$$\begin{array}{cccccccccccc}
 & & & & & & 0 & 1 & 0 & 1 & 1 & 1 \\
 & & & & & & & -1 & & +2 & & -2 \\
\hline
1 & 1 & 1 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 & 0 & & \\
1_1 & 1_1 & 1_1 & 0_2 & 1_1 & 0_1 & 0_1 & 1 & & & & \\
\hline
1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0
\end{array}$$

$$\begin{array}{r} 110011 \\ \times\,101100 \\ \hline \end{array}$$

$$\begin{array}{cccccccccccc}
 & & & & & & 1 & 1 & 0 & 0 & 1 & 1 \\
 & & & & & & & -1 & & -1 & & 0 \\
\hline
 & & & & & & & & & & & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & & \\
0 & 0 & 0 & 0_1 & 1_1 & 1_1 & 0_1 & 1 & & & & \\
\hline
0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0
\end{array}$$

$$\begin{array}{r} 110101 \\ \times\,011011 \\ \hline \end{array}$$

$$\begin{array}{cccccccccccc}
 & & & & & & 1 & 1 & 0 & 1 & 0 & 1 \\
 & & & & & & & +2 & & -1 & & -1 \\
\hline
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & & \\
1 & 1 & 1 & 0 & 1 & 0_1 & 1 & {}_1 & & & & \\
\hline
1 & 1 & 1 & 0 & 1 & 1 & 0 & 1 & 0 & 1 & 1 & 1
\end{array}$$

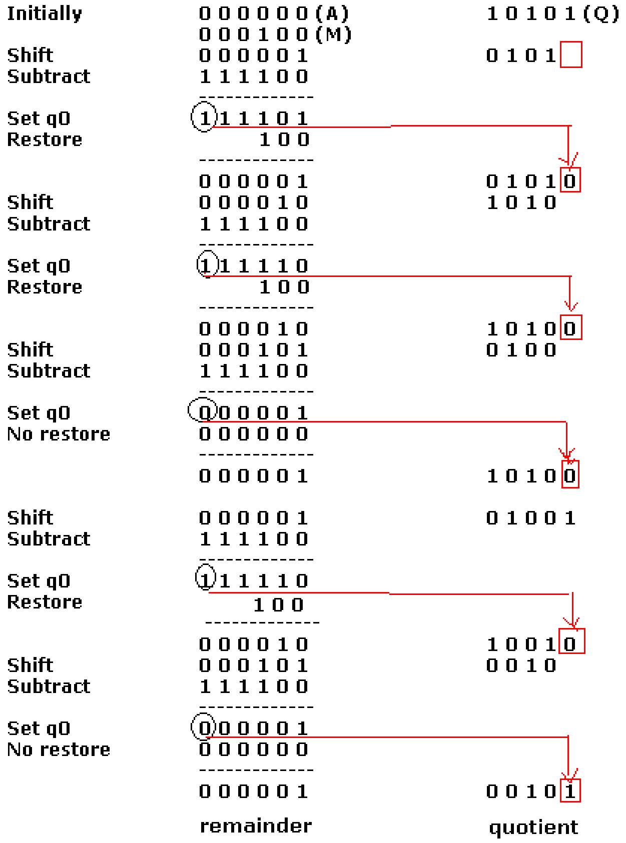
$$\begin{array}{r} 001111 \\ \times\,001111 \\ \hline \end{array}$$

$$\begin{array}{cccccccccccc}
 & & & & & & 0 & 0 & 1 & 1 & 1 & 1 \\
 & & & & & & & +1 & & & & -1 \\
\hline
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 1 & 1 & 1 & 1 & & & & \\
\hline
0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 1
\end{array}$$

Problem 6:

Given A=10101 and B=00100, perform A/B using restoring division algorithm.

Solution:



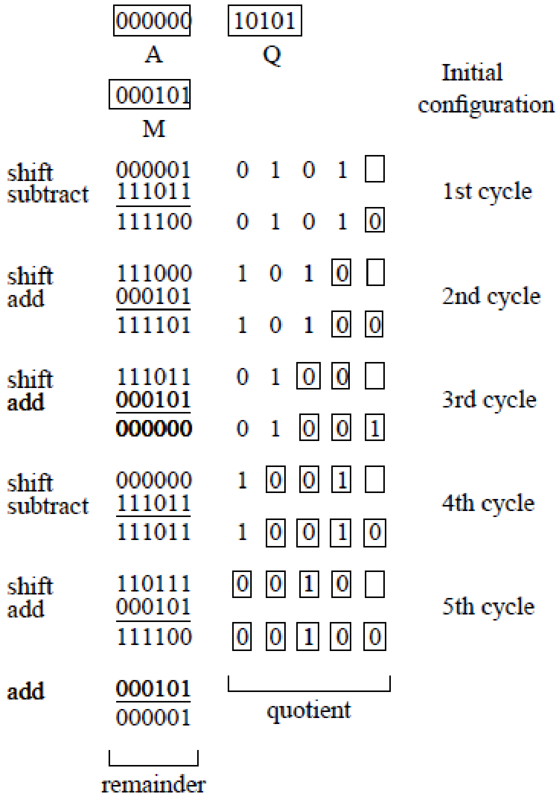
```
Initially      0 0 0 0 0 0 (A)        1 0 1 0 1 (Q)
               0 0 0 1 0 0 (M)
Shift          0 0 0 0 0 1            0 1 0 1 □
Subtract       1 1 1 1 0 0
               ------------
Set q0         1 1 1 1 0 1
Restore              1 0 0
               ------------
               0 0 0 0 0 1            0 1 0 1 0
Shift          0 0 0 0 1 0            1 0 1 0
Subtract       1 1 1 1 0 0
               ------------
Set q0         1 1 1 1 1 0
Restore              1 0 0
               ------------
               0 0 0 0 1 0            1 0 1 0 0
Shift          0 0 0 1 0 1            0 1 0 0
Subtract       1 1 1 1 0 0
               ------------
Set q0         0 0 0 0 0 1
No restore     0 0 0 0 0 0
               ------------
               0 0 0 0 0 1            1 0 1 0 0

Shift          0 0 0 0 0 1            0 1 0 0 1
Subtract       1 1 1 1 0 0
               ------------
Set q0         1 1 1 1 1 0
Restore             1 0 0
                ------------
               0 0 0 0 1 0            1 0 0 1 0
Shift          0 0 0 1 0 1            0 0 1 0
Subtract       1 1 1 1 0 0
               ------------
Set q0         0 0 0 0 0 1
No restore     0 0 0 0 0 0
               ------------
               0 0 0 0 0 1            0 0 1 0 1
               remainder              quotient
```

Problem 7:

Given A=10101 and B=00101, perform A/B using non-restoring division algorithm.

Solution:



	A	Q	
	$\boxed{000000}$	$\boxed{10101}$	Initial configuration
	$\boxed{000101}$ M		
shift	000001	0 1 0 1 □	1st cycle
subtract	$\underline{111011}$		
	111100	0 1 0 1 $\boxed{0}$	
shift	111000	1 0 1 $\boxed{0}$ □	2nd cycle
add	$\underline{000101}$		
	111101	1 0 1 $\boxed{0}$ $\boxed{0}$	
shift	111011	0 1 $\boxed{0}$ $\boxed{0}$ □	3rd cycle
add	$\underline{000101}$		
	000000	0 1 $\boxed{0}$ $\boxed{0}$ $\boxed{1}$	
shift	000000	1 $\boxed{0}$ $\boxed{0}$ $\boxed{1}$ □	4th cycle
subtract	$\underline{111011}$		
	111011	1 $\boxed{0}$ $\boxed{0}$ $\boxed{1}$ $\boxed{0}$	
shift	110111	$\boxed{0}$ $\boxed{0}$ $\boxed{1}$ $\boxed{0}$ □	5th cycle
add	$\underline{000101}$		
	111100	$\boxed{0}$ $\boxed{0}$ $\boxed{1}$ $\boxed{0}$ $\boxed{0}$	
add	$\underline{000101}$	quotient	
	000001 remainder		

Problem 8:

Represent 1259.12510 in single precision and double precision formats

Solution:

Step 1: Convert decimal number to binary format

$1259_{(10)} = 10011101011_{(2)}$

Fractional Part

$0.125_{(10)} = 0.001$

Binary number = 10011101011+0.001

= 10011101011.001

Step 2: Normalize the number

$10011101011.001 = 1.0011101011001 \times 2^{10}$

Step 3: Single precision format:

For a given number S=0, E=10 and M=0011101011001

Bias for single precision format is = 127

$E' = E + 127 = 10 + 127 = 137_{(10)}$

$= 10001001_{(2)}$

Number in single precision format is given as

$$\underbrace{0}_{\text{Sign}}\ \underbrace{10001001}_{\text{Exponent}}\ \underbrace{0011101011001\ldots0}_{\text{Mantissa(23 bit)}}$$

Step 4: Double precision format:

For a given number S=0, E=10 and M=0011101011001

Bias for double precision format is = 1023

$E' = E + 1023 = 10 + 1023 = 1033_{(10)}$

$= 10000001001_{(2)}$

Number in double precision format is given as

$$\underbrace{0}_{\text{Sign}}\ \underbrace{10001001}_{\text{Exponent}}\ \underbrace{0011101011001\ldots0}_{\text{Mantissa(23 bit)}}$$