ARITMÉTICA ENTERA

AMPLIACIÓN DE ESTRUCTURA DE COMPUTADORES

José Luis Risco, Carlos García, Eduardo Huedo Facultad de Informática

Aritmética

- 1.- Representación de reales
- 2.- Suma y resta de enteros
- 3.- Multiplicación de enteros

Multiplicación de enteros sin signo

Multiplicación de enteros con signo

Algoritmo de Booth

Multiplicadores recodificados

Salva-arrastre

Multiplicación combinacional

Árboles de Wallace

Multiplicador de Pezaris

Multiplicador de Baugh-Wooley

Multiplicadores recodificados

4.- División de enteros

División de enteros sin signo

División por convergencia

División combinacional

5.- Cálculos trigonométricos: método Cordic

Bibliografía

- 1.-"Computer arithmetic algorithms". I. Koren, Prentice Hall, 2002
- 2.- "Computer Arithmetic: Algoritms and Hardware Design", B.Parhami. Oxford UP. 2000
- 3.-"Computer architecture. A quantitative approach". Hennessy & Patterson, Morgan Kaufmann, 1995. Apéndice A.
- 4.-"Computer arithmetic". K. Hwang, John Wiley & Sons, 1979.
- 5.- "Digital computer arithmetic", J. Cavanagh, McGraw Hill, 1985

Representación de reales

• Representación en coma fija

- Número fijo de bits para la parte entera y para la parte decimal. Dejando implícito que la coma decimal se coloca entre ellos.
- La aritmética se realizará procesando la parte entera y la decimal separadamente usando instrucciones aritméticas enteras.
- Rango de representación bastante limitado.

• Representación en coma flotante

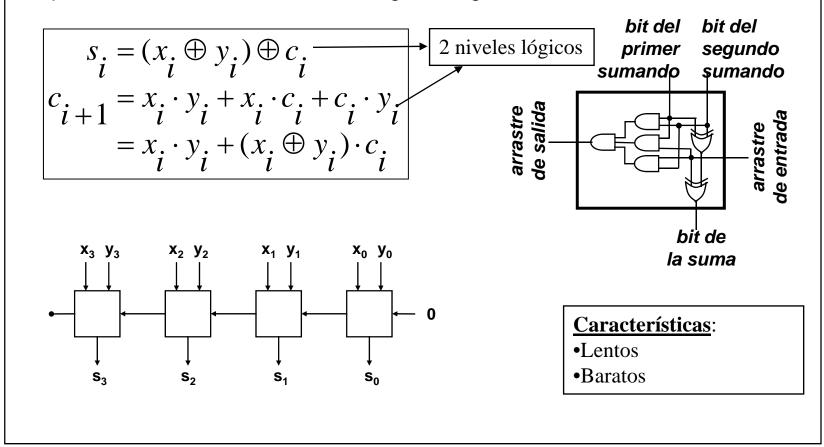
- Las aplicaciones científicas frecuentemente usan números muy pequeños o muy grandes y con la representación en coma fija estos números sólo tienen cabida si se utilizan muchos bits para la representación.
- Esta notación consta de los siguientes elementos:
 - s = signo
 - F = fracción
 - E = exponente
 - R = raiz
- El valor V se calculará como: $V = (-1)^{s*}F^*R^E$.

Suma y resta de enteros

- Sean *x* e *y* dos números enteros, representados por los vectores de bits X e Y. El algoritmo de la suma, SUMA, produce como resultado el vector de dígitos S que representa al número entero *s*, tal que s=x+y.
 - S=SUMA(X,Y)
- Para la resta, introducimos la operación cambio de signo (CS), siendo D un vector que representa al resultado *d* de la resta, d=x-y,
 - D=SUMA(X,CS(Y)).
- Si el rango de enteros representable por S o D es el mismo que el de X o Y, el resultado de la suma o resta puede no ser representable. En ese caso el resultado del algoritmo no sería correcto y se debería dar una señal de error por rebose.

Sumador con propagación de arrastres

- •Diseñar una celda combinacional que, tomando dos dígitos y un posible arrastre anterior, genere la suma y un arrastre posterior
- •Replicar la celda tantas veces como dígitos tenga el número

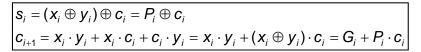


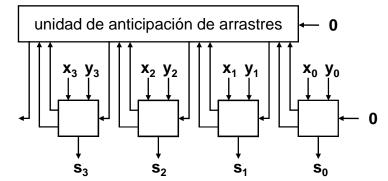
Sumador con anticipación de arrastres

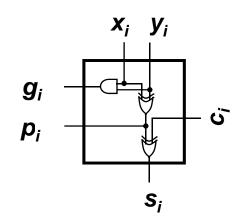
•Se busca reducir el tiempo de cálculo:

$$c_{i+1} = x_i \cdot y_i + (x_i \oplus y_i) \cdot c_i = g_i + p_i c_i$$

- •g_i es 1 si una celda genera un arrastre, es decir x_i=y_i=1
- $ullet p_i$ es 1 si una celda propaga un arrastre, es decir $x_i \oplus y_i = 1$







 $C_{2} = G_{1} + P_{1} \cdot C_{1}$ $= G_{1} + P_{1} \cdot G_{0} + P_{1} \cdot P_{0} \cdot C_{0}$ $C_{3} = G_{2} + P_{2} \cdot C_{2}$ $= G_{2} + P_{2} \cdot G_{1} + P_{2} \cdot P_{1} \cdot G_{0} + P_{2} \cdot P_{1} \cdot P_{0} \cdot C_{0}$ $C_{4} = G_{3} + P_{3} \cdot C_{3}$ $= G_{3} + P_{3} \cdot G_{2} + P_{3} \cdot P_{2} \cdot G_{1} + P_{3} \cdot P_{2} \cdot P_{1} \cdot G_{0} + P_{3} \cdot P_{2} \cdot P_{1} \cdot P_{0} \cdot C_{0}$

 $\mathbf{C}_1 = \mathbf{G}_0 + \mathbf{P}_0 \cdot \mathbf{C}_0$

- •4 niveles de puertas para cualquier bit de la suma
- •Conforme aumenta el número de bits el número de términos producto y el número de factores en ellos crece demasiado: para 32 bits el cálculo de c_{32} tiene 33 t.p y 33 factores

Sumador con varios niveles de anticipación de arrastres

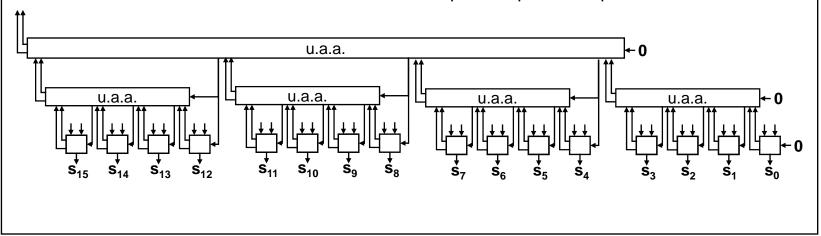
- •Utilizar la anticipación de arrastres, no solamente sobre celdas de un bit de suma, sino sobre módulos de suma con mayor número de bits.
 - •Un módulo genera un arrastre si se genera en alguna de sus celdas internas y se propaga hasta la salida
 - •Un módulo propaga un arrastre si el arrastre de entrada es uno y todas las celdas intermedias lo propagan

Para 4 bits, dichas señales resultan:

$$G^* = G_3 + G_2 \cdot P_3 + G_1 \cdot P_2 \cdot P_3 + G_0 \cdot P_1 \cdot P_2 \cdot P_3$$

$$P^* = P_0 \cdot P_1 \cdot P_3 \cdot P_4$$

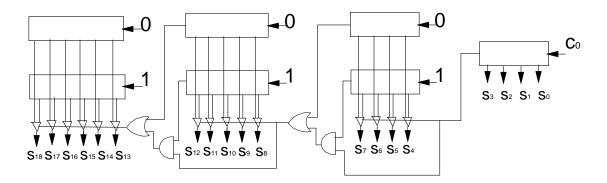
- •El arrastre de salida se calcula como: $c_{out} = G^* + P^* \cdot c_{in}$
- •Cada nuevo nivel en cascada añade 2 niveles de puertas para cualquier bit de la suma



Sumador con selección de arrastres

Idea:

- •Dividir el sumador en módulos que se implementen con alguno de los métodos anteriores.
- •Duplicar el nº de módulos de forma que se calculen en paralelo los resultados para $c_i = 0$ y para $c_i = 1$.
- •Cuando se conoce el c_i real se selecciona el valor adecuado.

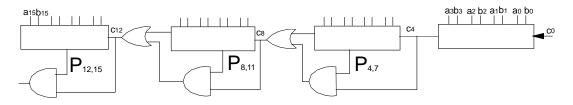


Los niveles de puertas requeridos para realizar una suma vienen dados por el camino crítico entre el arrastre de entrada y la selección del bit más significativo de la suma

Sumador con puenteo de arrastres (carry-skìp adder)

Idea:

- Aprovechar los datos de generación y propagación de arrastres sin usar un módulo de anticipación de arrastres.
 - $\bullet P_{03} = p_0 p_1 p_2 p_3$
 - $\bullet G_{03} = g_3 + p_3 g_2 + p_3 p_2 g_1 + p_3 p_2 p_1 g_0$
 - ➤ Modificamos ligeramente los sumadores para poder calcular.
 - La señal G es el arrastre de salida de cada módulo si el arrastre de entrada era 0
- •Es una mezcla de propagación y generación.



•Los niveles de puertas requeridos para realizar una suma vienen dados por el camino crítico entre el arrastre de entrada y el cálculo del bit más significativo de la suma

Sumador carry save

Objetivo:

- •Acelerar la suma cuando se tienen que sumar mas de dos operandos, tratando de evitar la propagación de arrastres.
- •Al sumar dos operandos los arrastres no se propagan, sino que se usan como tercer operando en la suma siguiente. Sólo en la última suma habrá que propagar los arrastres.
- •Ejemplo en decimal: S=357+409+143+112

Con propagación de arrastres

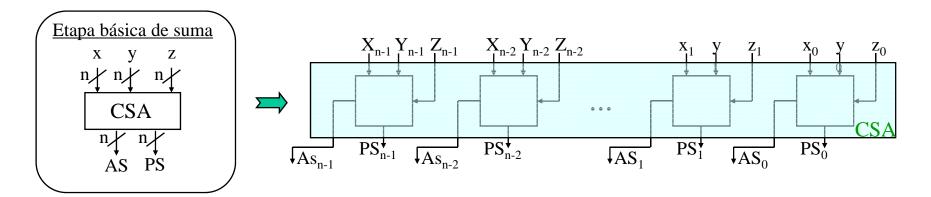
$$357
+409
766
+143
909
+112
1021$$

Con salva-arrastre

$$\begin{array}{c}
357 \\
409 \\
CSA + \underline{143} \\
899 \leftarrow PseudoSuma (PS) \\
001 \leftarrow Arrastre Salvado (AS) \\
CSA + \underline{112} \\
911 (PS) \\
Paso Final + \underline{11} (AS) \\
1021
\end{array}$$

Sumador carry save

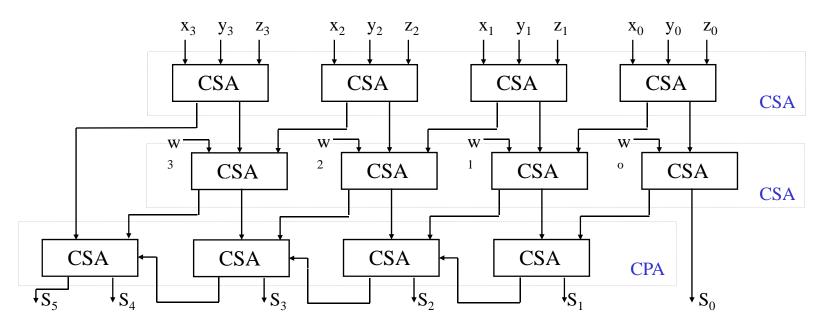
Construcción de un sumador salva-arrastre



Cada una de las celdas individuales del CSA son sumadores completos, en los que el arrastre se utiliza como tercer sumando.

Sumador carry save

Ej: diseño de un sumador carry save de 4 operandos de 4 bits.

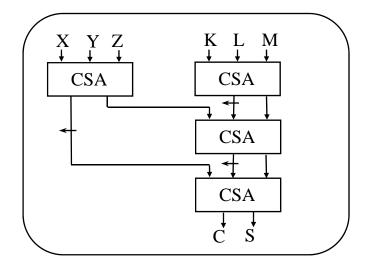


CSA: Carry Save Adder CPA: Carry Propagate Adder

Sumador carry save: Árboles de Wallace

Es otra forma de organizar los CSA para tratar de mejorar el rendimiento.

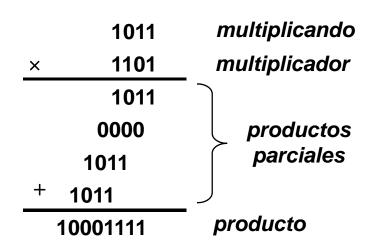
Arbol de Wallace con 6 operandos



Número de niveles en un árbol de Wallace para k operandos

Nº de operandos	Nº de niveles
3	1
4	2
5≤k≤6	3
7≤k≤9	4
10≤k≤13	5
14≤k≤19	6
20≤k≤28	7
29≤k≤42	8
43≤k≤63	9

Multiplicación de enteros sin signo



Método tradicional de multiplicación :

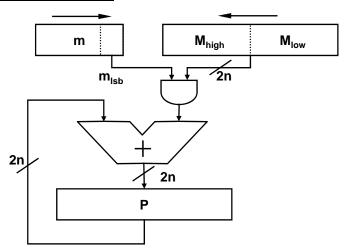
- Obtener los productos parciales
- Cada producto parcial debe estar desplazado una posición a la izquierda respecto al producto parcial anterior
- Una vez calculados todos los productos parciales se suman para obtener el producto
- Para un multiplicando de n bits y un multiplicador de m bits el producto ocupa como máximo n+m bits

Ruta de datos:

- 3 registros: multiplicador y producto
- 1 sumador de 2 entradas, en cada iteración sumar el producto parcial obtenido a la suma de los anteriores
- Para alinear correctamente los productos parciales, en cada iteración desplazar el multiplicando a la izquierda
- Para leer del mismo lugar cada uno de los bits del multiplicador, en cada iteración desplazarlo a la derecha

Multiplicación de enteros sin signo

Ruta de datos:



Algoritmo:

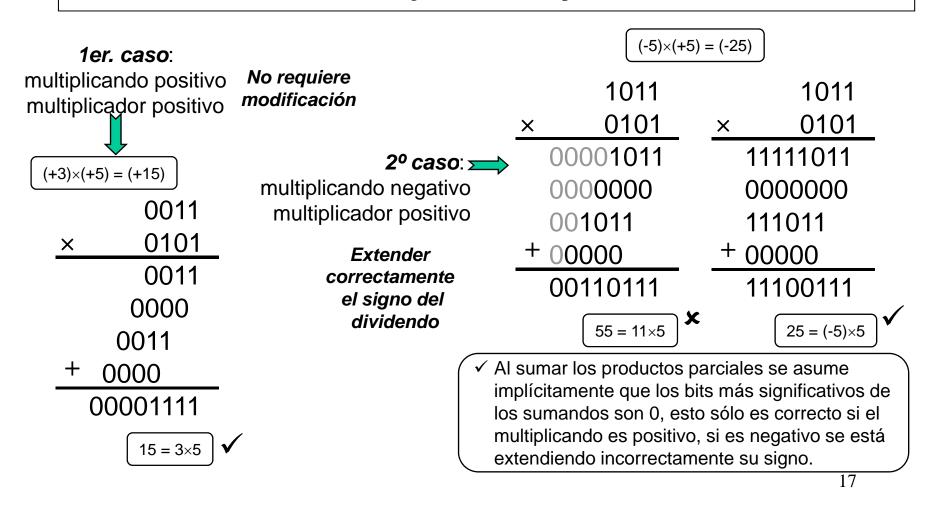
S0: cargar multiplicador en mcargar multiplicando en M_{low} borrar M_{high} borrar PS1: si m_{lsb} = 1 entonces $P \leftarrow P + M$ si m_{lsb} = 0 entonces $P \leftarrow P + 0$ S2: desplazar M a la izquierda

desplazar *m* a la derecha si *S1-S2* no se han repetido *n* veces, ir a *S1*

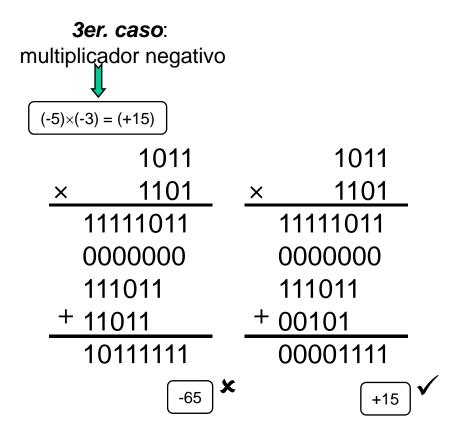
tras	m	М	P
S0	110 <u>1</u>	0000 1011	00000000
S1	110 1	00001011	00001011
S2	011 <u>0</u>	00010110	00001011
S1	0110	00010110	00001011
S2	001 <u>1</u>	00101100	00001011
S1	0011	00101100	00110111
S2	000 <u>1</u>	01011000	00110111
S1	0001	01011000	10001111
S2	0000	10110000	10001111

Multiplicación de enteros con signo

La multiplicación en C2 **no es coherente** con la multiplicación sin signo. Sin embargo, puede modificarse el algoritmo de suma y desplazamiento para que opere directamente en C2. Para ello distinguiremos 3 casos posibles:



Multiplicación de enteros con signo



En la última iteración sumar o restar el multiplicando en función del signo del multiplicador

Estudiemos la operación que se realiza cuando el **multiplicador es negativo**:

✓ El multiplicador está representado en C2 luego su valor es:

$$V(m) = -m_{n-1} \cdot 2^{n-1} + \sum_{i=0}^{n-2} m_i \cdot 2^i$$

✓ Si se ignora el signo del multiplicador (y se realiza la multiplicación teniendo en cuenta sólo los n-1 bits menos significativos del multiplicador) el resultado del método tradicional es:

$$V(P) = V(M) \cdot \sum_{i=0}^{n-2} m_i \cdot 2^i$$

$$= V(M) \cdot (V(m) + m_{n-1} \cdot 2^{n-1})$$

$$= V(M) \cdot V(m) + V(M) \cdot m_{n-1} \cdot 2^{n-1}$$

✓ El resultado buscado es V(M)·V(m), luego cuando m_{n-1}= 1 es necesario corregir el resultado parcial obtenido, restando el multiplicando a la mitad mas significativa de P:

$$V(M)\cdot V(m)=V(P)-V(M)\cdot m_{n-1}\cdot 2^{n-1}$$

- Permite multiplicar directamente enteros representados en C2.
- Evita ejecutar sumas consecutivas cuando el multiplicador presenta cadenas de *0s* o de *1s*.

<u>Idea</u>: Convertir el <u>multiplicador</u> en un número recodificado sobre un sistema binario no canónico bajo la forma de dígitos con signo:

Sistema binario canónico $D=\{0,1\}$ => Sistema binario no canónico $D=\{-1,0,1\}$.

Dada la cadena de bits:

teniendo en cuenta la igualdad:

$$2^{i+k-1} + 2^{i+k-2} + ... + 2^i = (2^{k-1} + 2^{k-2} + ... + 2^0)2^i = (2^k-1)2^i = 2^{i+k-2}i$$

la cadena de 1's podemos sustituirla por:

Peso: ...
$$i+k$$
, $i+k-1$, $i+k-2$, ..., $i+1$, i , $i-1$...

Bits: 1 0 0 ... 0 -1 0

(k-1) 0's seguidos

Para el caso de números negativos, en C2, cuyo bit más significativo, i+k es un 1, podemos demostrar también que se cumple esta equivalencia:

Peso: ...
$$i+k$$
, $i+k-1$, $i+k-2$, ..., $i+1$, i , $i-1$...

Bits: $\underbrace{1 \quad 1 \quad \dots \quad 1 \quad 1}_{k+1} \quad 0$

Esto es equivalente a:

$$\begin{array}{l} -2^{i+k} + 2^{i+k-1} + 2^{i+k-2} + ... + 2^i = -2^{i+k} + (2^{k-1} + 2^{k-2} + ... + 2^0) 2^i = \\ = -2^{i+k} + (2^k-1) \ 2^i = -2^{i+k} + 2^{i+k} - 2^i = -2^i \end{array}$$

Es decir, un -1 en la posición i.

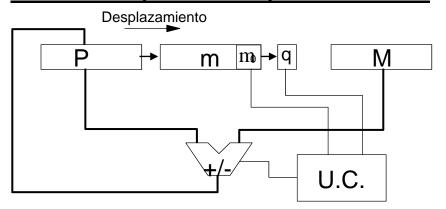
Recodificación del multiplicador: $(Y_{n-1}, Y_{n-2}, ..., Y_0)$:

- Se analizan todos los bits del multiplicador y se sustituyen las cadenas de 1s por un (-1) en la posición del 1 menos significativo y un 1 en la posición del 0 que las precede.
- De este modo resultará un multiplicador representado con los dígitos {-1,0,1} y al realizar la multiplicación habrá que hacer **sumas y restas**.

En resumen:

Bits del multiplicador	Dígito recodificado	Operación a realizar
$Y_i Y_{i-1}$	Z_{i}	
0 0	0	0 * multiplicando
0 1	1	1 * multiplicando
1 0	-1	-1 * multiplicando
1 1	0	0 * multiplicando

Ruta de datos para un multiplicador de Booth



Algoritmo de multiplicación

1. M \leftarrow multiplicando | m \leftarrow multiplicador | P = q \leftarrow 0.

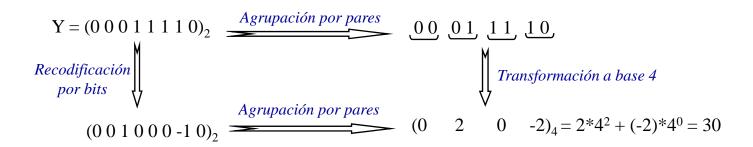
$$\begin{array}{lll} \textbf{2.} & \text{Si } m_o q = & 00 \text{ o } m_o q = & 11 & = & \text{DD}_{arit}(P,m,q) \\ & \text{Si } m_o q = & 10 & = & \text{P} \leftarrow P - M \mid DD_{arit}(P,m,q) \\ & \text{Si } m_o q = & 01 & = & \text{P} \leftarrow P + M \mid DD_{arit}(P,m,q) \\ \end{array}$$

El paso 2 se realiza n veces, siendo n el número de bits del multiplicador. Una vez finalizado, el resultado se hallará en los registros \mathbf{P} (parte más significativa) y \mathbf{m} (parte menos significativa).

Multiplicadores recodificados: Recodificación por pares de bits

- Como método de aceleración de la multiplicación se puede proceder tratando en cada paso un grupo de 2 bits del multiplicador en vez de uno solo.
- Nos servirá para multiplicar directamente números en C2, garantizando que para un multiplicador de n bits habrá como máximo n/2 productos parciales.
- Se realiza la misma acción que en la recodificación de Booth pero considerando pares de bits, junto con el bit que está a su derecha.

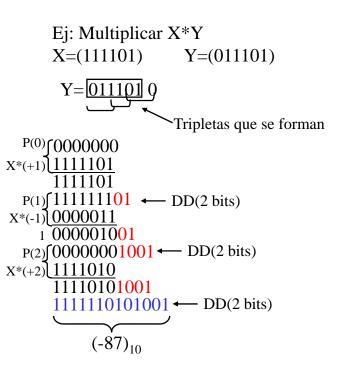
Ejemplo:



Multiplicadores recodificados: Recodificación por pares de bits

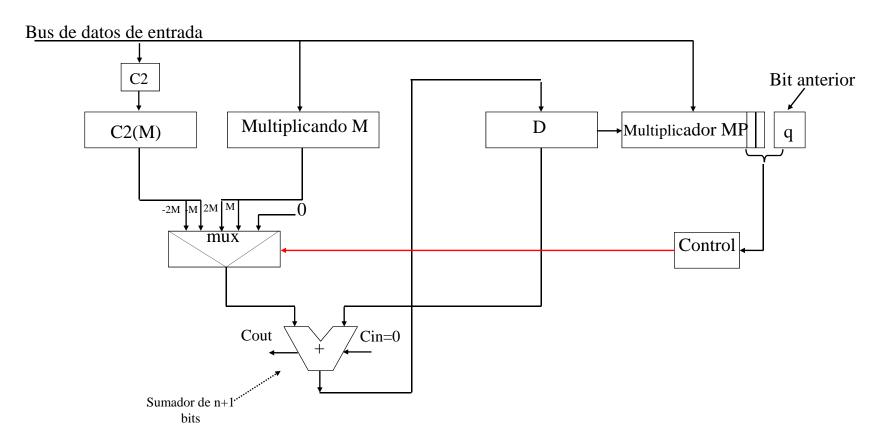
• El proceso de recodificación por pares de bits puede hacerse de forma directa, observando cada par de bits del multiplicador y el bit a su derecha.

Par d	le bits	Bit anterior	Digito base 4 recodificado		Operación a realizar
i+1	i	i-1			
0	0	0	(0 0)	0	0* multiplicando
0	0	1	(0 1)	1	+1* multiplicando
0	1	0	(1 - 1)	1	+1* multiplicando
0	1	1	(10)	2	+2* multiplicando
1	0	0	(-1 0)	-2	-2* multiplicando
1	0	1	(-1 1)	-1	-1* multiplicando
1	1	0	(0-1)	-1	-1* multiplicando
1	1	1	(00)	0	0* multiplicando



Multiplicadores recodificados: Recodificación por pares de bits

Recodificación por pares de bits: ruta de datos:



Multiplicación salva-arrastre

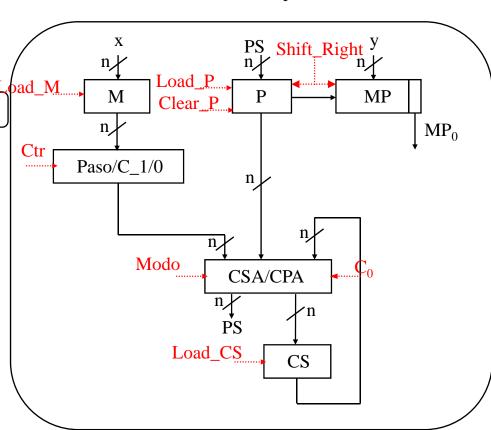
<u>Idea</u>: Usar sumadores salva-arrastre, dado que hay que realizar sumas con varios sumandos

Cada paso implica:

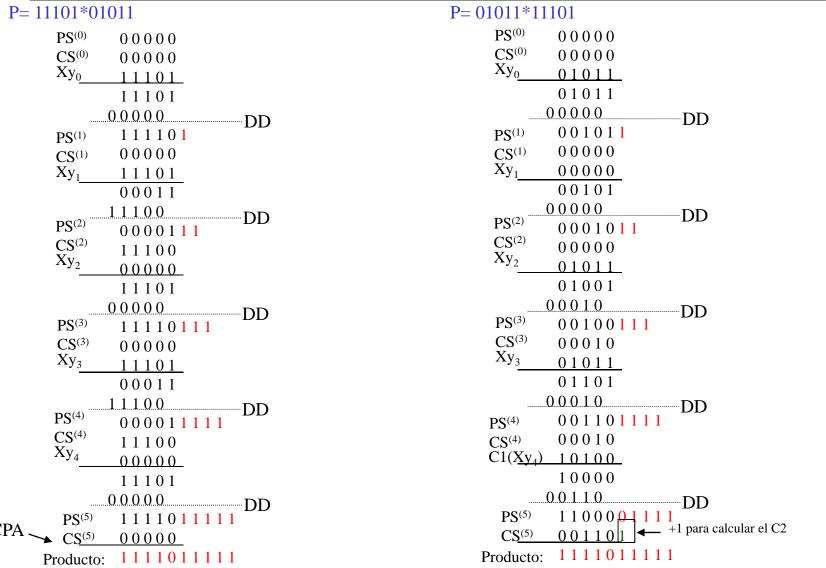
$$(PS^{(j+1)},\,CS^{(j+1)}) \mathrel{<=} SR(CSA(PS^{(j)},\,CS^{(j)},\,X^*Y_i))$$

CSA/CPA= Carry save adder/Carry propagate adder

Ruta de datos de un multiplicador CS



Multiplicación salva-arrastre

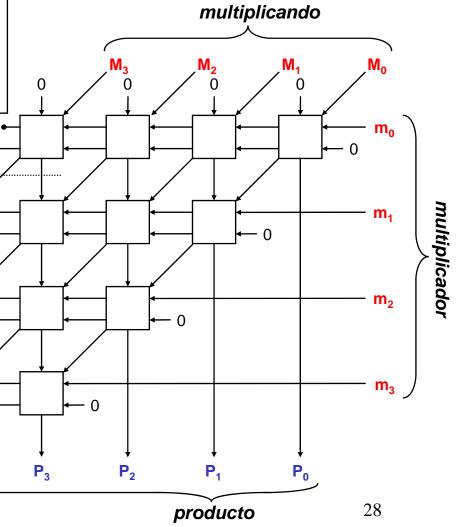


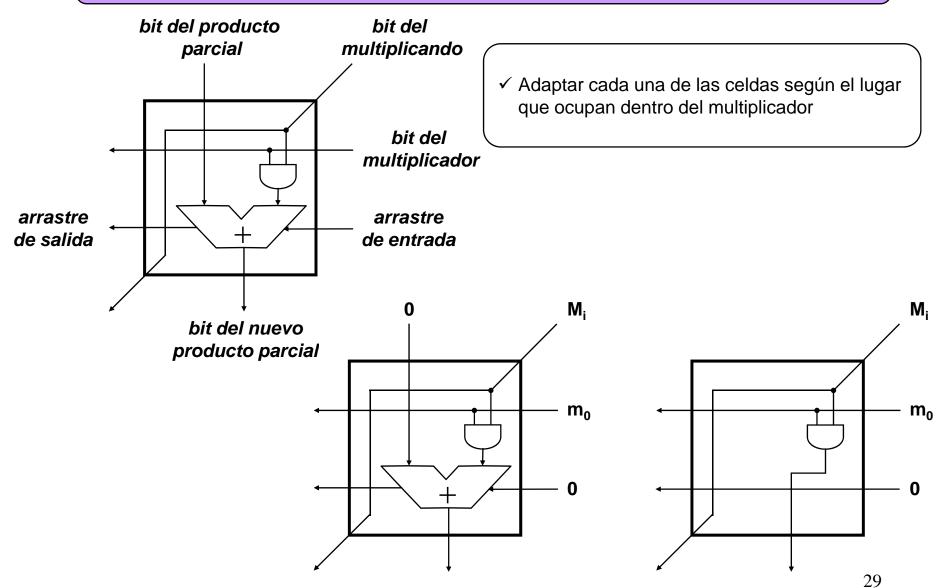
 P_5

 P_4

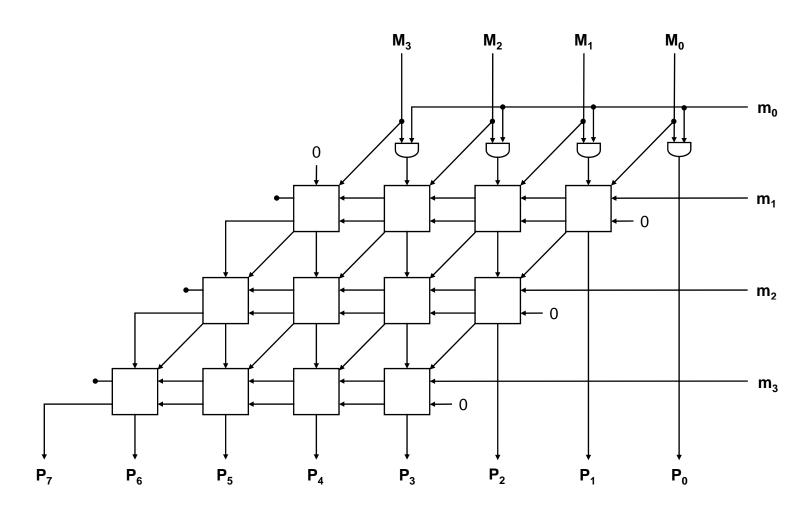
 Diseñar una celda combinacional que, tomando un dígito del multiplicando, otro del multiplicador y otro del producto parcial anterior, genere un bit del siguiente producto parcial

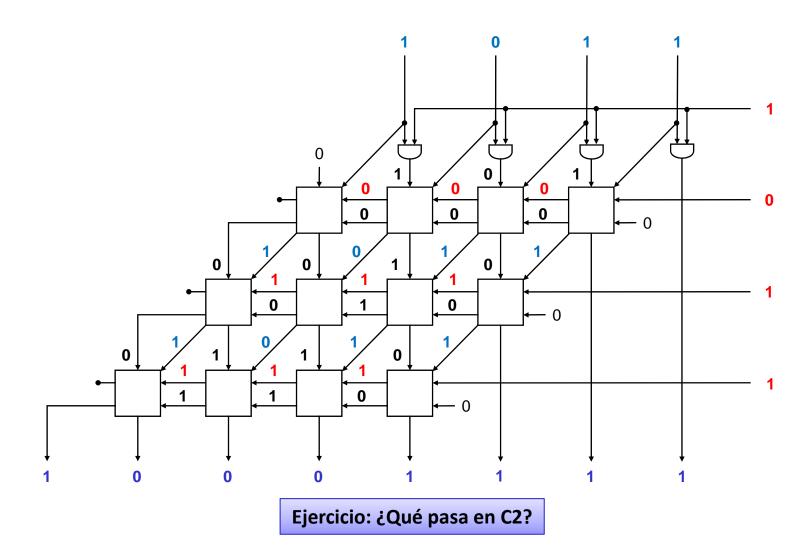
 Replicar la celda tantas veces como bits parciales haya que generar





Multiplicador combinacional (4 bits)





Multiplicación directa en C2: Multiplicador de Pezaris

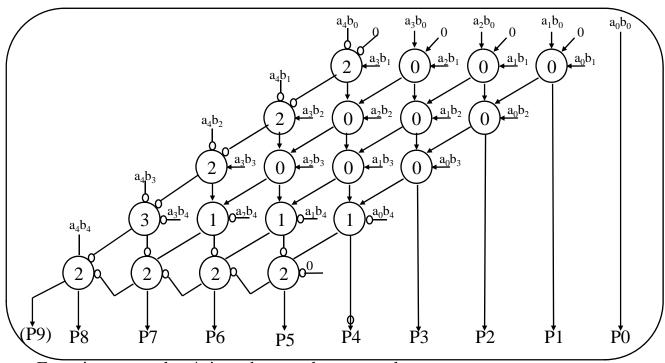
Un n° m(m_{n-1}, m_{n-2},... m₁, m₀) representado en C2 tiene un valor igual a: $V(m) = -m_{n-1} \cdot 2^{n-1} + \sum_{i=0}^{n-2} m_i \cdot 2^i$

La multiplicación de dos números $A=(a_{n-1}, a_{n-2}, ... a_1, a_0)$ y $B=(b_{n-1}, b_{n-2}, ... b_1, b_0)$ representados en C2 será por tanto:

$$A * B = A * \sum_{i=0}^{n-2} b_i * 2^i - A * b_{n-1} * 2^{n-1} = A * B = (\sum_{j=0}^{n-2} a_j * 2^j - a_{n-1} * 2^{n-1})(\sum_{i=0}^{n-2} b_i * 2^i) - (\sum_{j=0}^{n-2} a_j * 2^j - a_{n-1} * 2^{n-1}) * b_{n-1} * 2^{n-1} = A * B = \sum_{i=0}^{n-2} (\sum_{j=0}^{n-2} a_j * b_i * 2^{i+j}) - \sum_{i=0}^{n-2} a_{n-1} * b_i * 2^{n-1+i} - \sum_{j=0}^{n-2} a_j * b_{n-1} * 2^{j+n-1} + a_{n-1} * b_{n-1} * 2^{2n-2}$$

Entre paréntesis aparecen los productos con peso negativo

Multiplicador de Pezaris



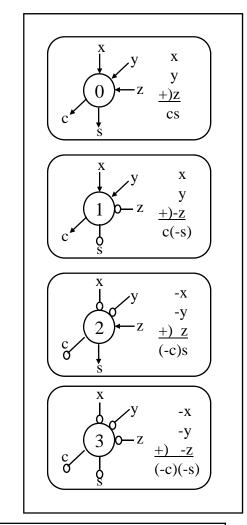
Ecuaciones para los 4 tipos de sumadores completos:

Tipos
$$0,1,2,3 \implies S = X \otimes Y \otimes Z$$

Tipos 0, 3
$$\Longrightarrow$$
 $C = XY + YZ + ZX$

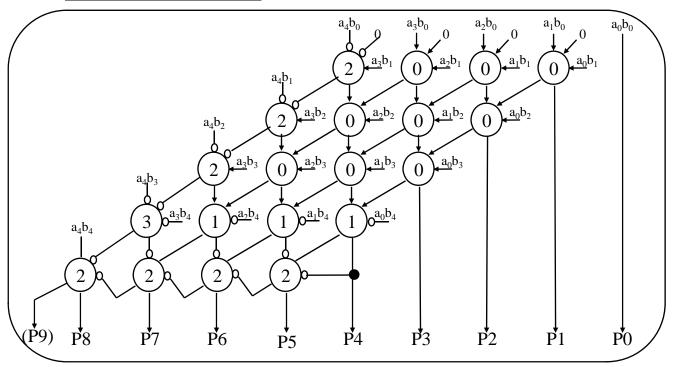
Tipo 1
$$\Longrightarrow C = XY + X\overline{Z} + Y\overline{Z}$$

Tipo 2
$$\Longrightarrow C = ZY + Z\overline{X} + Y\overline{X}$$



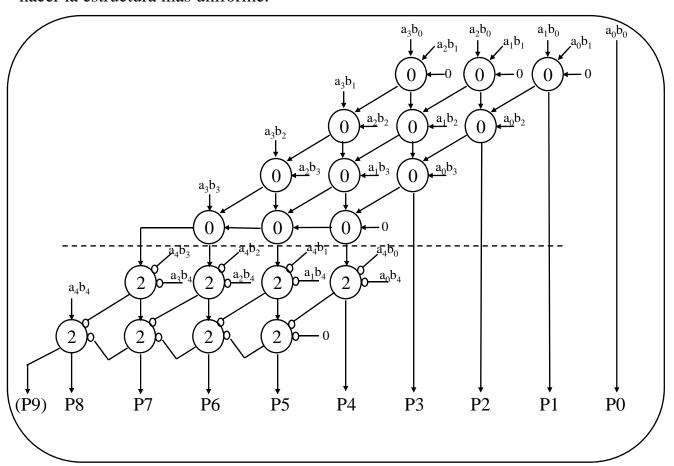
Nota: El peso de P4 es negativo. Para evitar este tipo de salidas podemos conectar p4 con la entrada derecha del sumador siguiente, como se muestra en la siguiente página.

Multiplicador de Pezaris



Variación del Multiplicador de Pezaris

Se trata de separar los sumandos positivos de los negativos, para reducir los tipos de sumadores, y hacer la estructura más uniforme.



Multiplicación directa en C2: Multiplicador de Baugh-Wooley

•Se busca incrementar la regularidad de la estructura usando sólo un tipo de sumador para todas las operaciones.

Fundamento:

• Se puede comprobar que <u>restar</u> el siguiente vector de m+1 bits:

$$X=(0,0, a_{m-2} k, a_{m-3} k, ..., a_0 k), k \in \{0,1\}$$

• Es igual a <u>sumar</u> los vectores Y y Z siendo:

$$Y = (0, \overline{k}, \overline{a_{m-2}}k, \overline{a_{m-3}}k, ..., \overline{a_0}k)$$
$$Z = (1, 1, 0, 0, ..., 0, k)$$

Demostración:

•Si k=0 => V(X)=0
V(Y+Z)=?
$$Y = (0,1,0,...,0)$$

 $Z = (1,1,0,...,0)$
 $1(0,0,0,...,0)$

Si k=1 => X=(0,0,
$$a_{m-2}$$
,..., a_0)
$$Y = (0,0,\overline{a_{m-2}},\overline{a_{m-3}},...,\overline{a_0})$$

$$Z = (1,1,0,0,...,1)$$

$$Y + Z = S1 + S2 \quad \text{con}$$

$$S1 = (1,1,\overline{a_{m-2}},\overline{a_{m-3}},...,\overline{a_0})$$

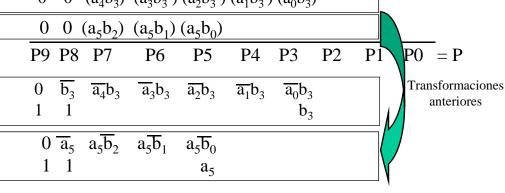
$$S2 = (0,0,0,0,...,0,1)$$
Por definición V(S1+S2)=V(C1(X)+1)=-X

Multiplicación combinacional

Multiplicación directa en C2: Multiplicador de Baugh-Wooley

A*B siendo A =
$$(a_5 a_4 a_3 a_2 a_1 a_0) m = 6$$

B = $(b_3 b_2 b_1 b_0) n = 4$

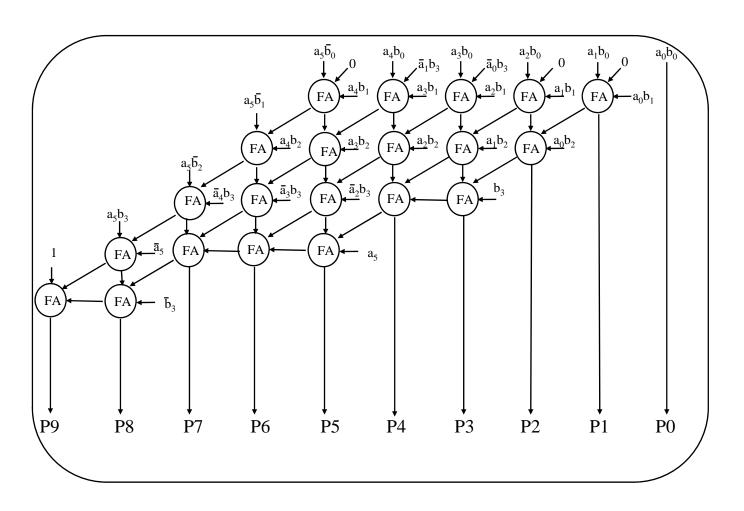


Entre paréntesis aparecen los productos negativos

$$\begin{cases} \overline{a_5} & \overline{a_4}b_3 & \overline{a_3}b_3 & \overline{a_2}b_3 & \overline{a_1}b_3 & \overline{a_0}b_3 \\ \overline{b_3} & a_5\overline{b_2} & a_5\overline{b_1} & a_5\overline{b_0} & \\ \underline{1 \quad 0} & a_5 & b_3 & \\ P9 \quad P8 \quad P7 \quad P6 \quad P5 \quad P4 & P3 \quad P2 \quad P1 \quad P0 \quad = P \end{cases}$$

Multiplicación combinacional

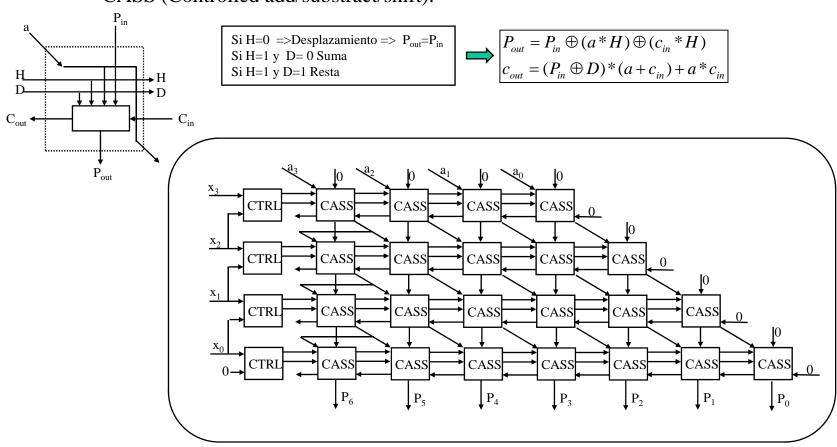
Multiplicador de Baugh-Wooley

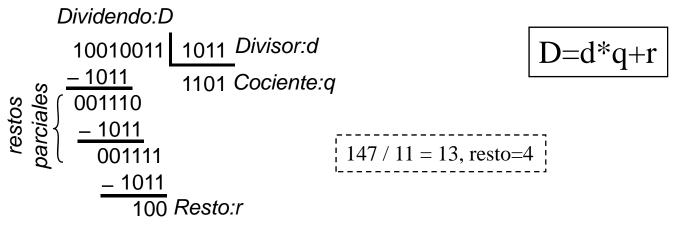


Multiplicación combinacional

Multiplicadores recodificados

- Basado en el algoritmo de Booth, pero combinacional.
- Se crea una celda básica que sume, reste o desplace en función de dos bits del multiplicador: CASS (Controlled add/substract/shift).





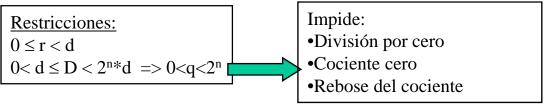
Suposiciones:

Dividendo=D => 2n bits

Divisor=d => n bits

Cociente: q => n bits

Resto: r => n bits

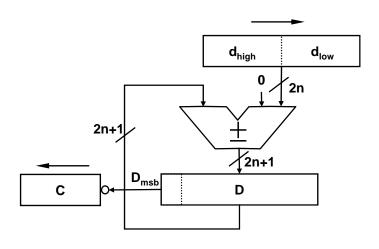


Método tradicional de división :

- Obtener los restos parciales y los bits del cociente recorriendo de izquierda a derecha los bits del dividendo:
 - si el resto parcial es mayor que el divisor, añadir un 1 al cociente; el nuevo resto parcial será la resta del resto parcial y del divisor
 - si el resto parcial es menor que el divisor, añadir un 0 al cociente y ampliar el resto parcial con un bit más del dividendo.

- usar 3 registros: resto/dividendo, divisor y cociente
- para alinear correctamente los restos parciales y el divisor, en cada iteración desplazar el divisor a la derecha
- para escribir en el mismo lugar cada uno de los bits del cociente, en cada iteración desplazarlo a la izquierda
- para evitar tener un comparador y un restador, usar éste último para comparar: el signo de la resta determinará si el resto parcial "cabe" entre el divisor

```
S0: cargar (0,dividendo) en D
cargar divisor en d_{high}
d_{low} = 0
C = 0
S1: D \leftarrow D - (0,d)
S2: si D_{msb} = 0 entonces
desplazar C a la izquierda insertando un 1
si D_{msb} = 1 entonces
desplazar C a la izquierda insertando un 0
D \leftarrow D + (0,d)
desplazar d a la derecha
si S1-S2 no se han repetido n+1 veces ir a S1
```



Ejemplo:

tras	D	d	С
S0	0 10010011	1011 0000	0000
S1	<u>1</u> 11100011	10110000	0000
S2	010010011	01011000	0000
S1	<u>0</u> 00111011	01011000	0000
S2	000111011	00101100	00 <mark>01</mark>
S1	<u>0</u> 00001111	00101100	0001
S2	000001111	00010110	0011
S1	<u>1</u> 11111001	00010110	0011
S2	000001111	00001011	0110
S1	<u>0</u> 00000100	00001011	0110
S2	00000 0100	00000101	1101

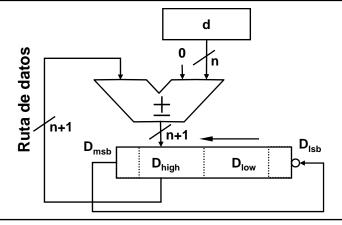
147	11*16=178	
-29		0 (0)
147	11*8 = 88	
59		1 (1)
59	11*4=44	
15		1 (3)
15	11*2=22	
-7		0 (6)
15	11	
4		1 (13)
4	11	13

División sin restauración:

- ✓ considerar la secuencia de operaciones que se realiza tras la resta en S2:
 - si D_{msb} = 0 ("cabe") se desplaza D a la izquierda y se resta d. Queda: 2-D-d
 - si D_{msb} = 1 ("no cabe") se suma d, se desplaza el resultado y se resta d. Queda: 2(D+d)-d = 2·D+d
- ✓ entonces, en lugar de restaurar:
 - sumar o restar d en función de D_{msh}

cargar (0 dividendo) en D

• en la última iteración restaurar el resto (sumándole d) si es necesario



0	o.	cargar (o,dividendo) en D
		cargar divisor en d
S	1:	D<<1
S	2:	$D_{high} \leftarrow D_{high} - (0,d)$
		si $D_{msb} = 0$ entonces $D_{lsb} \leftarrow 1$
		si $D_{msb} = 1$ entonces $D_{lsb} \leftarrow 0$
S	4:	si $D_{\text{msb}} = 0$: D<<1; $D_{\text{high}} \leftarrow D_{\text{high}} - (0,d)$
		si $D_{msb} = 1$: D<<1; $D_{high} \leftarrow D_{high} + (0,d)$
		si S3-S4 no se han repetido n-1 veces ir a S3
S	5:	si $D_{msb} = 0$ entonces $D_{lsb} \leftarrow 1$
		si $D_{msb} = 1$ entonces $D_{high} \leftarrow D_{high} + (0,d), D_{lsb} \leftarrow 0$
		mgn mgn (/ // /sb

tras	D	d
S0	0 1001 0011	1011
S1	10010 011 <i>0</i>	1011
S2	<u>0</u> 0111 011 <i>0</i>	1011
S3	00111 011 <u>1</u>	1011
S4	01110 11 <i>10</i>	1011
S4	<u>0</u> 0011 11 <i>10</i>	1011
S3	00011 11 <u>11</u>	1011
S4	<u>00111</u> 1 110	1011
S4	<u>1</u> 1100 1 <i>110</i>	1011
S3	11100 1 <u>110</u>	1011
S4	11001 1100	1011
S4	<u>0</u> 0100 <i>1100</i>	1011
S5	0 0100 <u>1101</u>	1011

1	(9, 3)
1	(18, 6) <
1	(7, 6) R
1	(7, 6) 1
1	(14, 12) <
1	(3, 12) R
1	(3, 12) 1
1	(7, 8) <
1	(-4, 8) R
1	(-4, 8) 0
1	(-7, 0) <
' 1	(4, 0) S
1	(4, 0) 1
<u> </u>	43

División por convergencia: Divisor multiplicativo

Se puede usar en sistemas que tengan un multiplicador rápido.

Sólo calcula el cociente: Q=A/B

Idea: Hallar una fracción equivalente a A/B pero con denominador 1.

Proceso iterativo:

$$B*F_0*F_1*...*F_n \to 1$$

 $A*F_0*F_1*...*F_n \to Q$

Suposiciones:

- •A y B son fracciones positivas
- •B está normalizado B=0,1xxxxx => B \geq 1/2
- •A está alineado convenientemente.

Proceso:

B= $1-\delta$ siendo $0 < \delta < \frac{1}{2}$

La secuencia de denominadores que se construyen son de la forma:

$$\bullet B_i = B_{i-1} * F_i$$

Eligiendo los F_i de forma que: $B_{i-1} < B_i < 1$

Puede tomarse: $\underline{F}_0 = 1 + \delta$

$$B_0 = B*F_0 = (1 - \overline{\delta})*(1 + \delta) = 1 - \delta^2$$

Con
$$F_0 > 1 \implies B_0 > B$$
, y $B_0 < 1$

En la siguiente iteración se elige: $F_1 = 1 + \delta^2$

$$B_1 = B_0 * F_1 = (1 - \delta^2) (1 + \delta^2) = 1 - \delta^4$$

Y de nuevo $B_1 > B_0$.

Para la i-ésima iteración: $F_i = 1 + \delta^{2^i}$

$$B_{i} = B_{i-1} * F_{i} = (1 - \delta^{2^{i}})(1 + \delta^{2^{i}}) = 1 - (\delta^{2^{i}})^{2} = 1 - \delta^{2^{i+1}}$$

Convergencia:

De la fórmula anterior se puede ver que B_i converge exponencialmente a 1.

También se puede ver teniendo en cuenta que:

$$\delta \le \frac{1}{2} => \delta^2 \le \frac{1}{4} => B_0 = 1 - \delta^2 \ge 0.11_{bin}$$

 $B_1 = 1 - \delta^4 \ge 1 - 1/16 = 15/16 = 0.1111$

En cada paso el número de 1s iniciales se duplica. Por ej. Para una máquina de 64 bits, obtener B= 0,111....111 sólo requiere 6 pasos.

Cálculo práctico de las iteraciones:

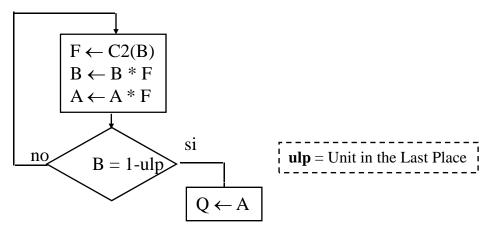
El cálculo de Fi es difícil y por ello habrá que calcularlo de otra manera. Se puede ver que:

$$F_i = 1 + \delta^{2^i} = 2 - (1 - \delta^{2^i}) = 2 - B_{i-1}$$

Y como $B_{i-1} = 0,1... => 2- B_{i-1} = C2(B_{i-1})$

Por tanto para calcular cada F_i solamente se necesita calcular el C2 de B_{i-1}.

Algoritmo:



División por convergencia: Cálculo del recíproco

Se puede usar en sistemas que tengan un multiplicador rápido.

Sólo calcula el cociente: Q=A/B

Idea: *Hallar Q multiplicando A**(1/B)

Suposiciones:

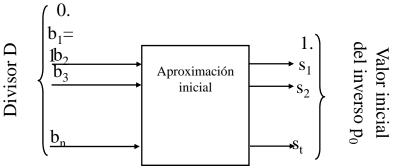
- •A y B son fracciones positivas
- •B está normalizado B=0,1xxxxx => 1>B≥1/2 => 1<1/B≤2
- •A<B para que Q<1.
- •A está alineado convenientemente.

Método:

Se parte de una aproximación inicial al valor del inverso de B, de la forma:

$$P_0 = 1, s_1 s_2 ... s_t$$

Esta aproximación inicial se puede almacenar en una tabla en ROM, de forma que dados los k bits más significativos de B (excluyendo el 0,1 inicial que es constante) la salida muestra una aproximación lo más exacta posible, sobre t bits al inverso de B.



Ej: Tabla que genera 4 bits del inverso a partir de 2 bits de B.

Entradas		Salidas (p_0)		
$b_2 b_3$	(valor decimal)	$s_1 s_2 s_3 s_4$	(Valor decimal)	
0 0 0 1 1 0 1 1	0,5 0,625 0,75 0,875	1 1 1 1 1 0 0 1 0 1 0 1 0 0 1 0	1,9375 1,5625 1,3125 1,125	

División por convergencia: Cálculo del recíproco

El proceso iterativo para calcular el inverso de B consiste en:

•p₀ se toma de la tabla.

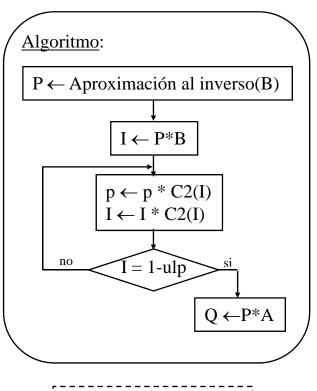
$$\bullet I_0 = p_0 * B$$

•
$$p_i = p_{i-1}^* (2 - I_{i-1})$$
• $I_i = I_{i-1}^* (2 - I_{i-1})$
Puede comprobarse que: $\lim_{i \to \infty} I_i = 1$
Por tanto: (1)

$$\begin{array}{ll} p_0 & I_0 = p_0^* B \\ p_1 = p_0^* (2 - I_0) & I_1 = I_0 (2 - I_0) = p_0^* B^* (2 - I_0) = p_1^* B \\ p_2 = p_1^* (2 - I_1) & I_2 = I_1 (2 - I_1) = p_1^* B^* (2 - I_1) = p_2^* B \end{array}$$

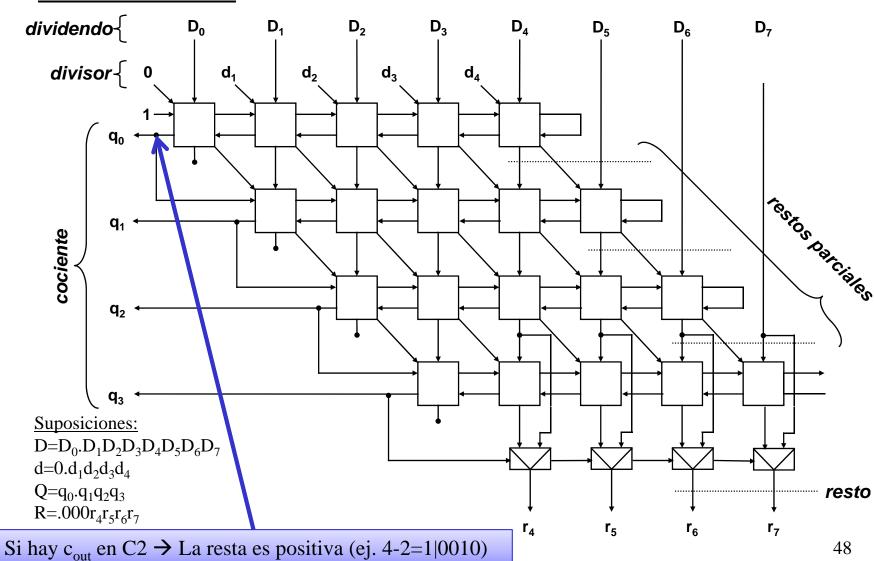
En general:

$$I_i = p_i *B \text{ y según (1): } \lim_{i \to \infty} p_i = \frac{1}{B}$$

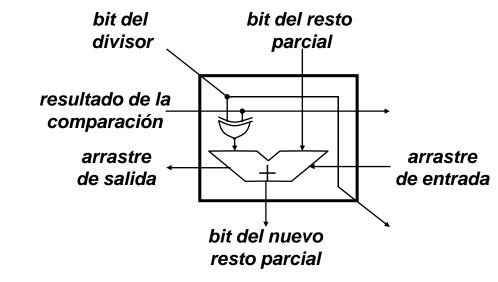


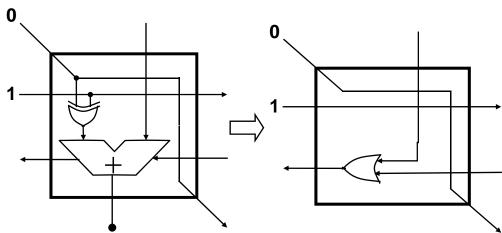
ulp = Unit in the Last Place

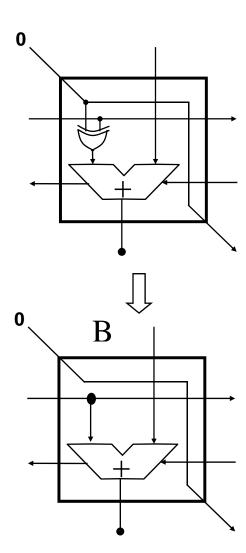
División sin restauración

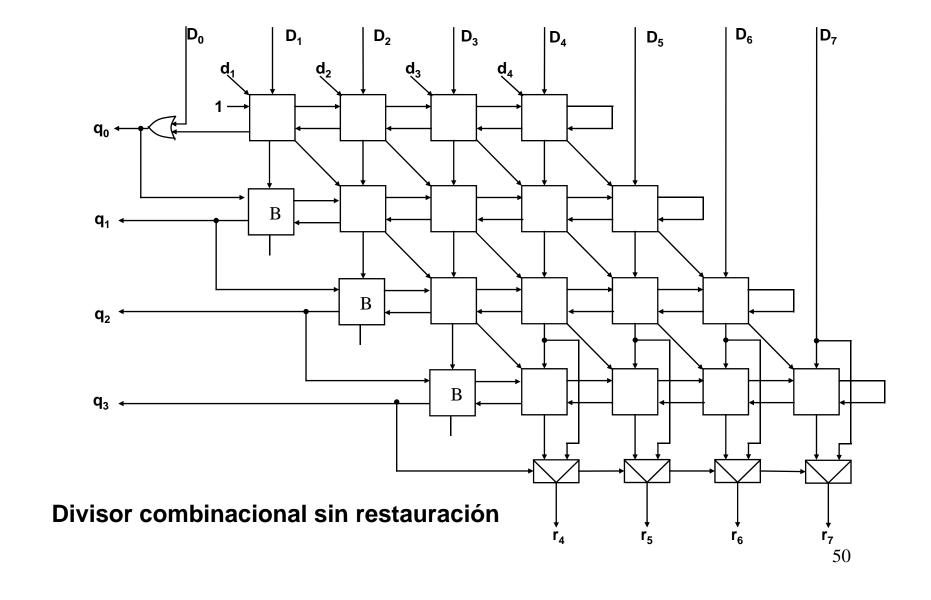


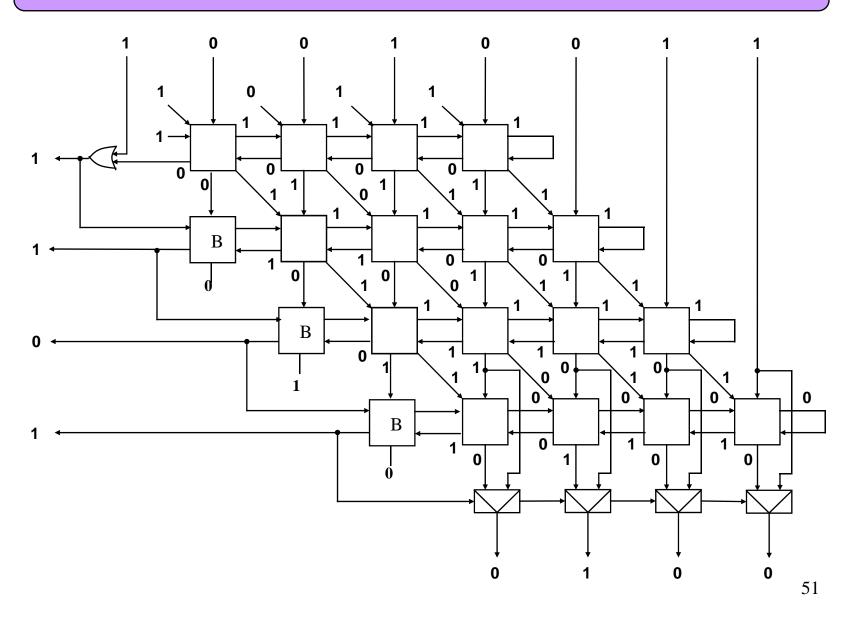
Celdas básicas







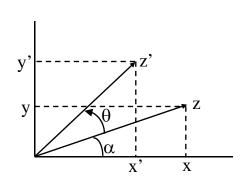




Cálculos trigonométricos: CORDIC

COordinate Rotation DIgital Computer

Objetivo: Calcular el seno de un ángulo de modo iterativo y sencillo



$$z=(x,y)$$

$$x = |z|\cos \alpha$$

$$y = |z|\sin \alpha$$

$$Z'=(x',y')$$

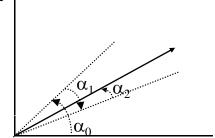
 $X' = |z'|\cos(\alpha + \theta) = |z|(\cos\alpha\cos\theta - \sin\alpha\sin\theta) = x\cos\theta - y\sin\theta$ Si la rotación hubiese sido horaria cambiaría el signo de y sen θ

$$X' = x \cos \theta - y \sin \theta = \cos \theta (x - y \tan \theta)$$

 $y' = y \cos \theta + x \sin \theta = \cos \theta (y + x \tan \theta)$

Si tan θ es una potencia de 2 => las coordenadas de z' pueden calcularse mediante sumas/restas y desplazamientos.

Partiendo de un vector z_0 de argumento 0^0 , hacer rotaciones sucesivas hasta alcanzar el ángulo θ , del que queremos hallar el seno o el coseno.



Cálculos trigonométricos: CORDIC

Proceso iterativo:

Se parte de : $z_0 = (x_0, y_0) = (x_0, 0)$

Al hacer sucesivas rotaciones, llegaremos a vectores z_{i+1} con coordenadas:

$$x_{i+1} = x_i \pm y_i * 2^{-i}$$

 $y_{i+1} = y_i \pm x_i * 2^{-i}$ El cálculo sólo implica
sumas/restas y desplazamientos

Después de n iteraciones se dispondrá de un vector z_n tal que:

$$|z_{n}| = \frac{1}{\cos \alpha_{n-1}} |z_{n-1}| = \frac{1}{\cos \alpha_{n-1}} \frac{1}{\cos \alpha_{n-2}} |z_{n-2}| = \dots = \frac{1}{\prod_{i=0}^{n-1} \cos \alpha_{i}} |z_{0}|$$

Pero:
$$\frac{\lim}{n \to \infty} \left(\prod_{i=0}^{n-1} \cos \alpha_i \right) = 0,6073$$

Por ello, eligiendo z_0 =(0,6073,0), para n suficientemente grande se obtendrá un vector z_n con módulo 1, y con argumento θ , y por tanto:

 $x_n = \cos \theta$

 $y_n = sen \theta$