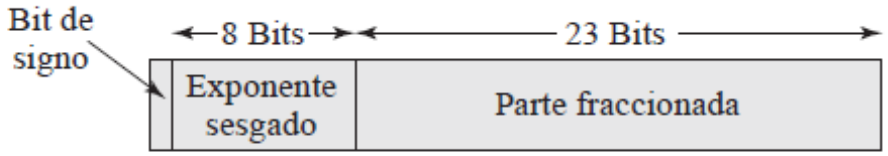
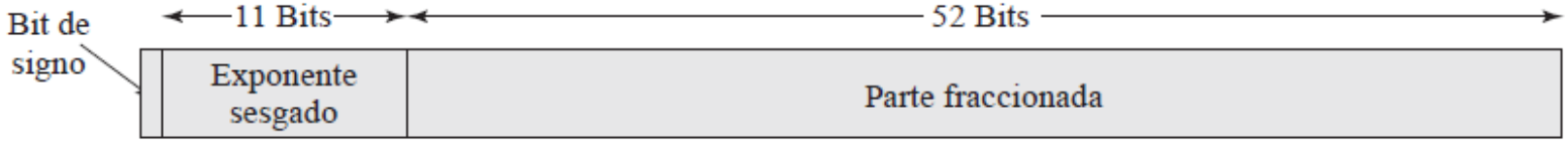
Aritmética en coma flotante

Estándares de 32 y 64 bits

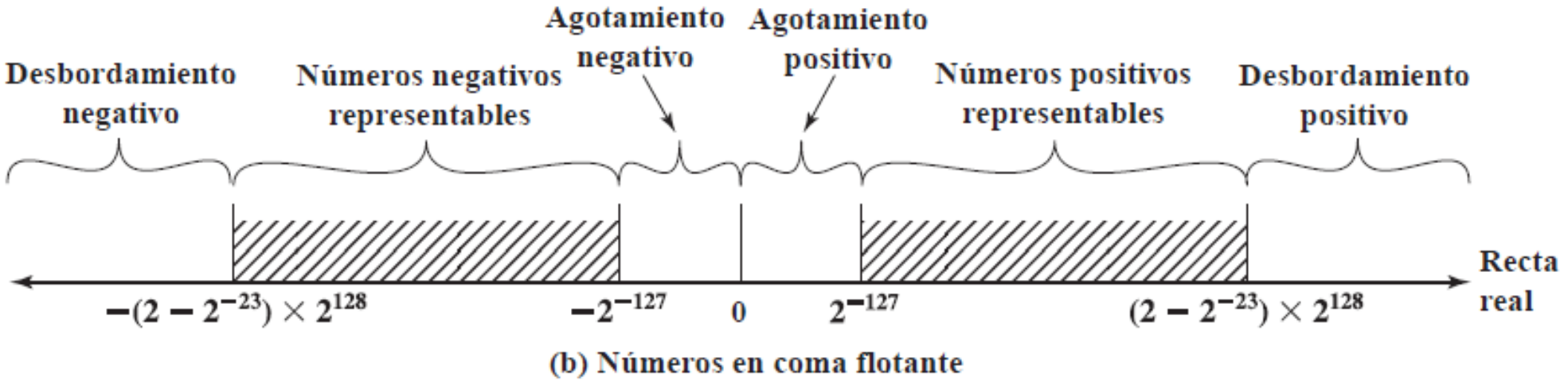


(a) Formato simple



(b) Formato doble

Intervalos representables por el estándar



(b) Números en coma flotante

Estándar IEEE 754 para coma flotante.

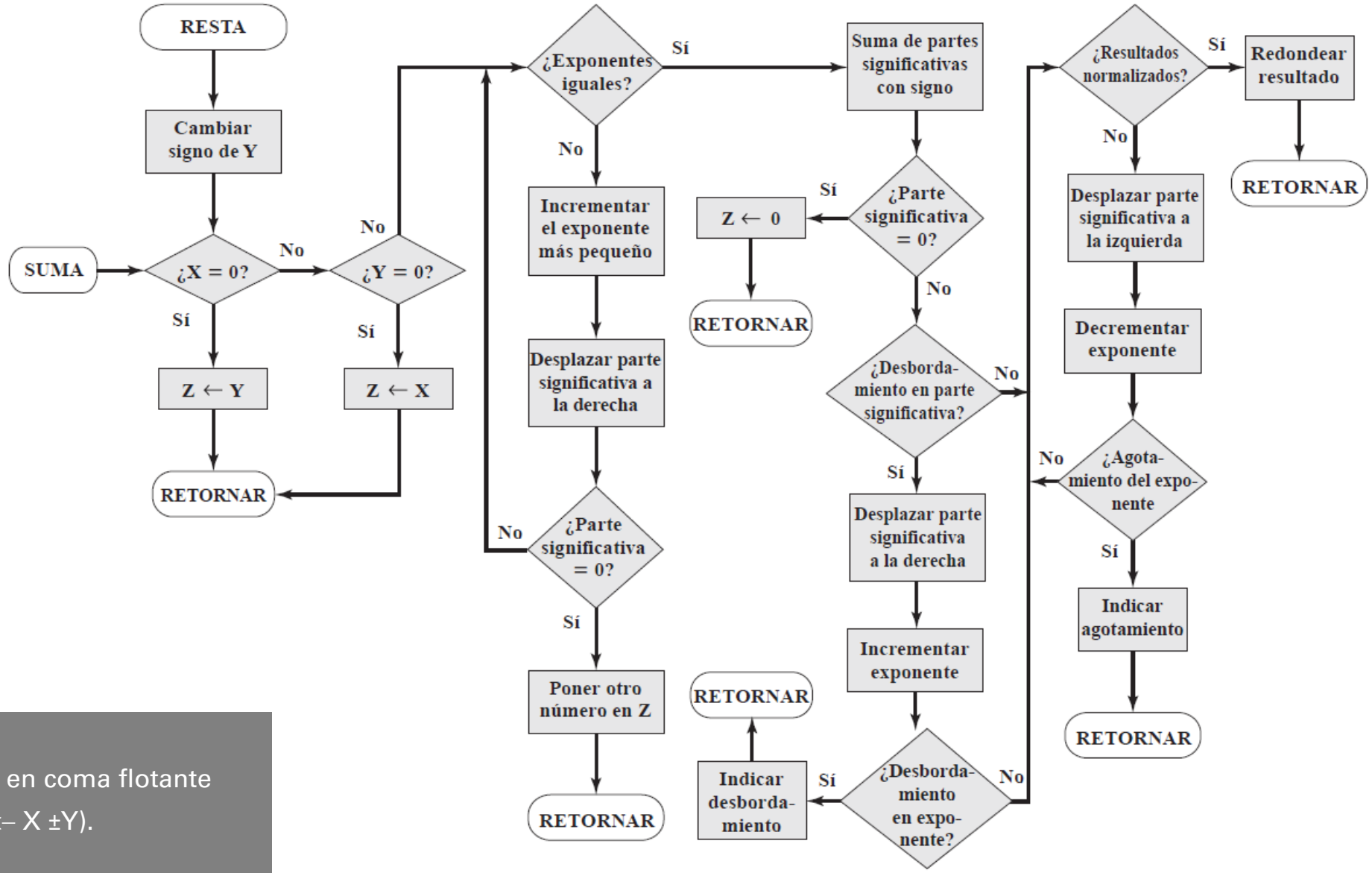
Parámetro	Formato			
	Simple	Simple ampliado	Doble	Doble ampliado
Longitud de palabra (bits)	32	≥43	64	≥79
Longitud de exponente (bits)	8	≥11	11	≥15
Sesgo del exponente	127	sin especificar	1023	sin especificar
Exponente máximo	127	≥1023	1023	≥16.383
Exponente mínimo	−126	≤−1022	−1022	≤−16.382
Rango de números (base 10)	10^{-38}, 10^{+38}	sin especificar	10^{-308}, 10^{+308}	sin especificar
Longitud de mantisa (bits)*	23	≥31	52	≥63
Número de exponentes	254	sin especificar	2046	sin especificar
Número de fracciones	2^{23}	sin especificar	2^{52}	sin especificar
Número de valores	$1{,}98 \times 2^{31}$	sin especificar	$1{,}99 \times 2^{53}$	sin especificar

	Precisión simple (32 bits)				Doble precisión (64bits)			
	Signo	**Exponente sesgado**	**Parte fraccionaria**	**Valor**	**Signo**	**Exponente sesgado**	**Parte fraccionaria**	**Valor**
Cero positivo	0	0	0	0	0	0	0	0
Cero negativo	1	0	0	−0	1	0	0	0
Más infinito	0	255 (todo unos)	0	∞	0	2047 (todo unos)	0	∞
Menos infinito	1	255)todo unos)	0	−∞	1	2047 (todo unos)	0	−∞
NaN silencioso	0 ó 1	255 (todo unos)	≠0	NaN	0 ó 1	2047 (todo unos)	≠0	NaN
NaN indicador	0 ó 1	255 (todo unos)	≠0	NaN	0 ó 1	1047 (todo unos)	≠0	NaN
Positivo normalizado ≠ cero	0	$0 < e < 255$	f	$2^{e-127}(1,f)$	0	$0 < e < 2047$	f	$2^{e-1023}(1,f)$
Negativo normalizado ≠ cero	1	$0 < e < 255$	f	$-2^{e-127}(1,f)$	1	$0 < e < 2047$	f	$-2^{e-1023}(1,f)$
Positivo denormalizado	0	0	f ≠ 0	$2^{e-126}(0,f)$	0	0	f ≠ 0	$2^{e-1022}(0,f)$
Negativo denormalizado	1	0	f ≠ 0	$-2^{e-126}(0,f)$	1	0	f ≠ 0	$-2^{e-1022}(0,f)$

Estándar IEEE 754 para coma flotante.

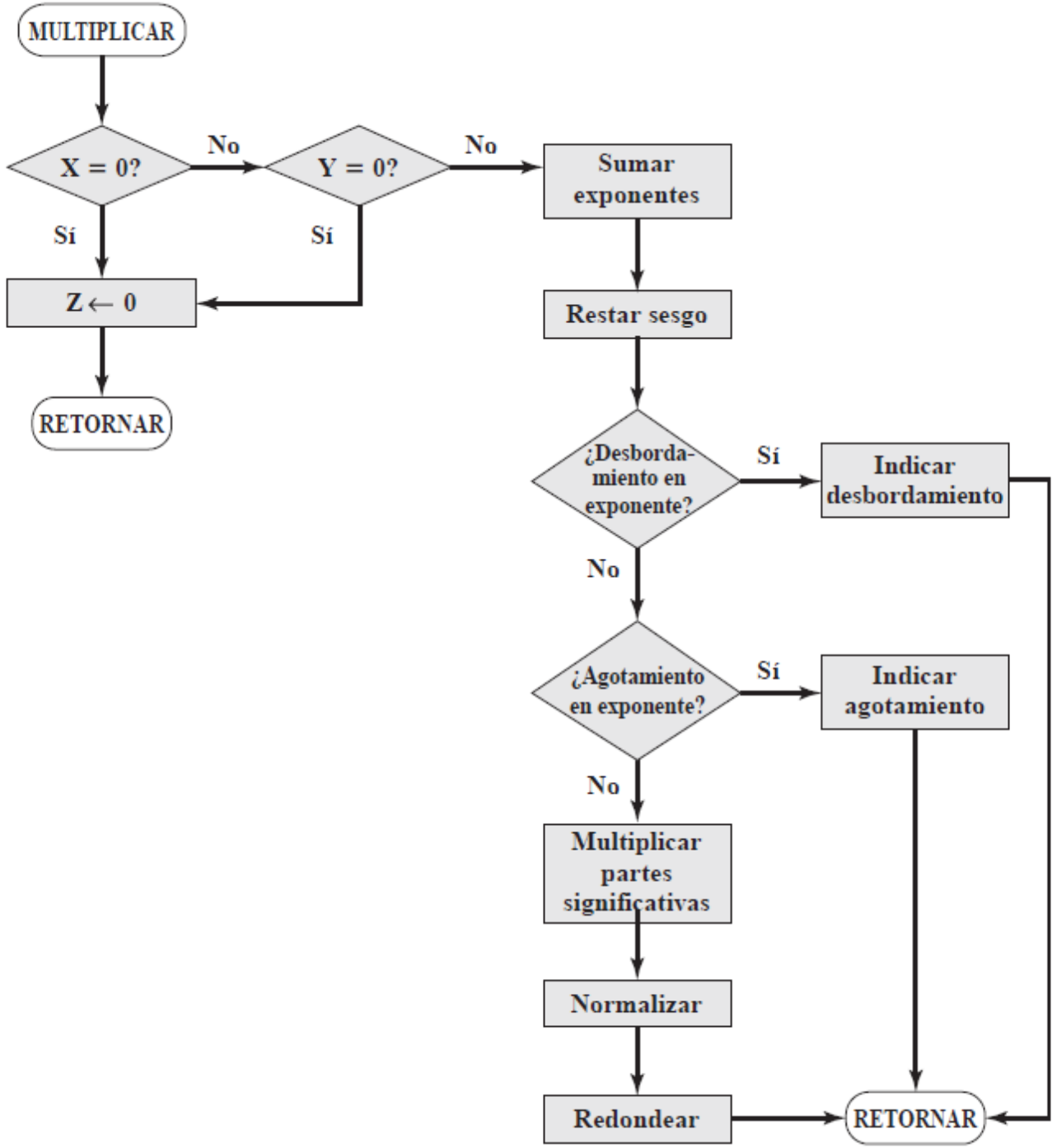
Artimética en coma flotante

Números en punto flotante	Operaciones aritméticas
$X = X_s \times B^{X_E}$ $Y = Y_s \times B^{Y_E}$	$\left.\begin{aligned} X + Y &= (X_s \times B^{X_E - Y_E} + Y_s) \times B^{Y_E} \\ X - Y &= (X_s \times B^{X_E - Y_E} - Y_s) \times B^{Y_E} \end{aligned}\right\} X_E \leq Y_E$ $X \times y = (X_S \times Y_S) \times B^{X_E + Y_E}$ $\frac{X}{Y} = \left(\frac{X_S}{Y_S}\right) \times B^{X_E - Y_E}$

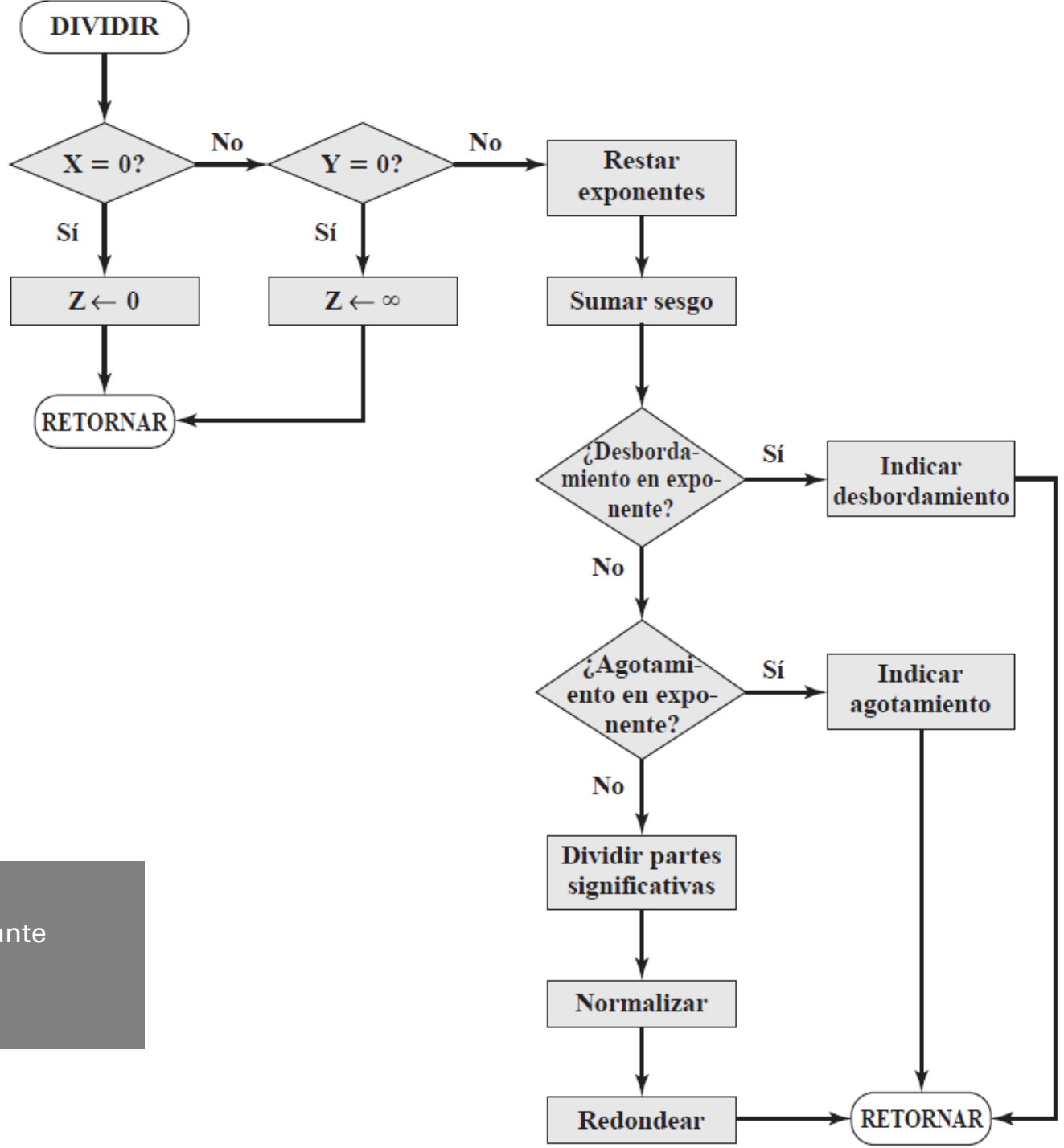


Suma y resta en coma flotante (Z <– X ±Y).

Multiplicación en coma flotante
(Z <– X ±Y).



División en coma flotante
(Z <– X ±Y).



Referencias

- Organización y Arquitectura. Stallings 7a 2005