Daha açık olarak, $\odot$ sembolü, dört temel aritmetik işlemi $+, -, *$, veya $\div$ den herhangi birini temsil etsin.
Eğer, x ve y makine sayıları olmak üzere $x \odot y$ hesaplanacak ve yüklenecek ise, bu durumda tek bir makine kelimesine uyacak şekilde $x \odot y$ yi yerleştirmek için yapabileceğimiz en iyi şey, onu $\mathrm{fl}(x \odot y)$ ye yuvarlamak ve o sayıyı yerleştirmektir.

Örnek: Yukarıdaki süreci açıklamak için, kayan-noktalı sayı sisteminde beş desimal nokta ile işlem yapan bir desimal makine kullanalım ve iki makine sayısı

$$x = 0.31426 \times 10^3 \qquad y = 0.92577 \times 10^5$$

in toplamında, farkında, çarpımında ve bölümündeki bağıl hataları belirleyelim.

Çözüm Ara sonuçlar için bir çift-uzunluklu işlemci kullanırsak

$$\begin{aligned} x+y &= 0.92891\,26000 \times 10^5 \\ x-y &= -0.92262\,74000 \times 10^5 \\ x*y &= 0.29093\,24802 \times 10^8 \\ x \div y &\approx 0.33945\,79647 \times 10^{-2} \end{aligned}$$

buluruz. Beş desimal noktalı bilgisayar, yuvarlanmış formda bunları

$$\begin{aligned} \mathrm{fl}(x+y) &= 0.92891 \times 10^5 \\ \mathrm{fl}(x-y) &= -0.92263 \times 10^5 \\ \mathrm{fl}(x*y) &= 0.29093 \times 10^8 \\ \mathrm{fl}(x \div y) &= 0.33946 \times 10^{-2} \end{aligned}$$

olarak yükler. Bu sonuçlardaki bağıl hatalar, sırası ile $2.8 \times 10^{-6}, 2.8 \times 10^{-6}, 8.5 \times 10^{-6}$, ve 6.0×10^{-6} olup, hepsi de 10^{-5} den küçüktür.

$$\mathrm{fl}(x) = x(1+\delta) \qquad |\delta| \leq 2^{-24}$$

Böylece,eğer x ve y makine sayıları ise,

$$\mathrm{fl}(x \odot y) = [x \odot y](1+\delta) \qquad |\delta| \leq 2^{-24}$$

Eğer x ve y nin makine sayıları olması gerekmiyorsa, karşılık gelen sonuç:

$$\mathrm{fl}(\mathrm{fl}(x) \odot \mathrm{fl}(y)) = (x(1+\delta_1) \odot y(1+\delta_2))\,(1+\delta_3) \qquad |\delta_i| \leq 2^{-24}$$

x, y ve z nin `Marc- 32` de makine sayıları olduklarını varsayalım ve $x(y+z)$ yi hesaplamak isteyelim.

$$\begin{aligned}
\mathrm{fl}[x(y+z)] &= [x\,\mathrm{fl}(y+z)](1+\delta_1) && |\delta_1| \leq 2^{-24} \\
&= [x(y+z)(1+\delta_2)](1+\delta_1) && |\delta_2| \leq 2^{-24} \\
&\approx x(y+z)(1+\delta_2+\delta_1+\delta_1\delta_2) && \\
&\approx x(y+z)(1+\delta_1+\delta_2) && \\
&= x(y+z)(1+\delta_3) && |\delta_3| \leq 2^{-23}
\end{aligned}$$

dir, çünkü $\delta_2\delta_1$ değeri 2^{-23} ile karşılaştırıldığında gözardı edilebilir.

Örnek (Duyarlılık Kaybı)

$$\begin{aligned} x &= 0.37214\ 78693 \\ y &= 0.37202\ 30572 \\ x - y &= 0.00012\ 48121 \end{aligned}$$

Eğer bu hesap beş-rakam mantissalı bir desimal bilgisayarda gerçekleştirilirse, bu durumda

$$\begin{aligned} \mathrm{fl}(x) &= 0.37215 \\ \mathrm{fl}(y) &= 0.37202 \\ \mathrm{fl}(x) - \mathrm{fl}(y) &= 0.00013 \end{aligned}$$

olur. Bu durumda, bağıl hata oldukça büyüktür:

$$\left| \frac{x - y - [\mathrm{fl}(x) - \mathrm{fl}(y)]}{x - y} \right| = \left| \frac{0.00012\ 48121 - 0.00013}{0.00012\ 48121} \right| \approx \%4$$

Kararlı-Kararsız Hesaplamalar

Eğer bir sayısal sürecin bir adımında yapılan küçük hatalar ardışık adımlarda büyüyorsa ve hesaplamanın tamanındaki duyarlılığı ciddi olarak azaltıyorsa, bu sayısal süreç **kararsızdır**.

Örnek

Reel sayıların, ardışık olarak

$$\begin{cases} x_0 = 1 \qquad x_1 = \frac{1}{3} \\ x_{n+1} = \frac{13}{3}x_n - \frac{4}{3}x_{n-1} \qquad (n \geq 1) \end{cases} \tag{1}$$

ile tanımlı dizisini göz önüne alalım. Bu indirgeme bağıntısının

$$x_n = \left(\frac{1}{3}\right)^n \tag{2}$$

dizisini oluşturduğu kolayca görülebilir.

Örnek

$$
\begin{array}{ll}
x_0 = 1.00000\,00 & \\
x_1 = 0.33333\,33 & (\text{yuv. 7 rak. doğru}) \\
x_2 = 0.11111\,12 & (\text{yuv. 6 rak. doğru}) \\
x_3 = 0.03703\,73 & (\text{yuv. 5 rak. doğru}) \\
x_4 = 0.01234\,56 & (\text{yuv. 4 rak. doğru})
\end{array}
\qquad
\begin{array}{ll}
x_5 = 0.00411\,87 & (\text{yuv. 3 rak.} \\
x_6 = 0.00138\,57 & (\text{yuv. 2 rak.} \\
x_7 = 0.00051\,31 & (\text{yuv. 1 rak.} \\
x_8 = 0.00037\,57 & (\text{yuv. 0 rak.} \\
x_9 = 0.00094\,37 &
\end{array}
$$

$$
\begin{aligned}
x_{10} &= 0.00358\,87 \\
x_{11} &= 0.01429\,27 \\
x_{12} &= 0.05715\,02 \\
x_{13} &= 0.22859\,39 \\
x_{14} &= 0.91437\,35 \\
x_{15} &= 3.65749\,3 \qquad (10^8 \text{ bağıl hata ile yanlış})
\end{aligned}
$$