国防科技大学学报
JOURNAL OF NATIONAL UNIVERSITY OF DEFENSE TECHNOLOGY
第18卷第2期 1996年6月 Vol. 18 No. 2

双字长浮点乘法运算算法设计与分析*

骆志刚
(国防科技大学研究生院 长沙 410073)
黄旭慧
(国防科技大学基础部 长沙 410073)

摘 要 文中给出了在计算机上实现双字长浮点乘法运算的算法的计算公式、计算步骤及误差估计，算法原理适用于一般计算机系统的任意字长浮点乘法运算。

关键词 计算机，数学子程序，算法
分类号 TP301.6

The Design and Analysis of an Algorithm for the Double—word Floating Point Multiplication

Luo Zhigang
(Graduate School, NUDT, Changsha, 410073)
Huang Xuhui
(Department of Basic Courses, NUDT, Changsha, 410073)

Abstract This essay presents an algorithm for the double-words floating point multiplication used in computer, and discusses the computation formula, the steps of computing and error estimation. The principle of this algorithm can be used in the floating point multiplication of any length of words in general computer system.

Key words computer, mathematics subroutine, algorithm

计算机系统，特别是面向大型科学和工程计算及大规模数据处理的巨型计算机系统，进行科学和工程计算时离不开调用各种数学子程序。双字长浮点乘法运算是最基本的数学子程序之一，其精度和速度对计算机系统双字长浮点运算的效率至关重要。

* 1995年9月22日收稿

1 计算公式

不妨设计算机的运算寄存器为 64 位，即一个单字长浮点数为 64 位，第 0 位表数符，第 1 至 15 位表阶码，第 16 至 63 位表尾数。

(1) 双字长数的浮点表示形式

源操作数 $X = x_f \cdot 2^{xj} \times 0.x_1x_2\cdots\cdots x_{96}$

$$= x_f \cdot 2^{xj} \times (X_1 \times 2^{-24} + X_2 \times 2^{-48} + X_3 \times 2^{-72} + X_4 \times 2^{-96}) \quad (1)$$

源操作数 $Y = y_f \cdot 2^{yj} \times 0.y_1y_2\cdots\cdots y_{96}$

$$= y_f \cdot 2^{yj} \times (Y_1 \times 2^{-24} + Y_2 \times 2^{-48} + Y_3 \times 2^{-72} + Y_4 \times 2^{-96}) \quad (2)$$

结 果 数 $Z = z_f \cdot 2^{zj} \times 0.z_1z_2\cdots\cdots z_{96}$

$$= z_f \cdot 2^{zj} \times (Z_1 \times 2^{-24} + Z_2 \times 2^{-48} + Z_3 \times 2^{-72} + Z_4 \times 2^{-96}) \quad (3)$$

$$= Z_f \cdot 2^{zj} \times ((Z_1Z_2) \times 2^{-48} + (Z_3Z_4) \times 2^{-96}) \quad (4)$$

其中 x_f，y_f，z_f 是数符；xj，yj，zj 是阶；x_i，y_i，z_i ($i=1, 2, \cdots, 96$) 是一位二进制的数字；X_i，Y_i，Z_i ($i=1, 2, 3, 4$) 是 24 位二进制数；Z_1Z_2，Z_3Z_4 是 48 位二进制数。

(2) 每个双字长浮点数用两个 64 位来表示。高部值（放在高字中）的格式与单字长浮点数格式相同，而整个 96 位尾数的低 48 位（称低部值）放在低字的低 48 位，低字的高 16 位必须是零，如图 1。

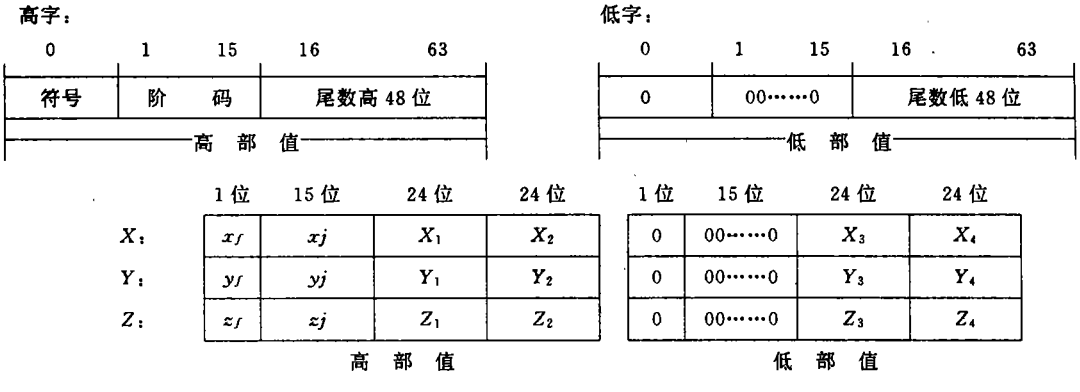

图 1

(3) 计算公式

由式 (1)、(2)、(3) 的浮点形式可得

$$Z^* = X * Y$$

$$= z_f \cdot 2^{zj} \times ((X_1 \times 2^{-24} + X_2 \times 2^{48} + X_3 \times 2^{-72} + X_4 \times 2^{-96}) \times (Y_1 \times 2^{-24} + Y_2 \times 2^{-48} + Y_3 \times 2^{-72} + Y_4 \times 2^{-96}))$$

$$= z_f \cdot 2^{zj} \times ((X_1 \times Y_1) \times 2^{-48} + (X_1 \times Y_2 + X_2 \times Y_1) \times 2^{-72} + (X_1 \times Y_3 + X_2 \times Y_2 + X_3 \times Y_1) \times 2^{-96} + (X_3 \times Y_2) \times 2^{-120} + (X_1X_2 \times Y_4 + X_4 \times Y_1Y_2 + X_2X_3 \times Y_3) \times 2^{-144} + (X_3 \times Y_4) \times 2^{-168} + (X_4 \times Y_3Y_4) \times 2^{-192}) \quad (5)$$

于是得到

$$Z = z_f \cdot 2^{zj} \times ([(X_1 \times Y_1) \times 2^{-48} + (X_1 \times Y_2 + X_2 \times Y_1) \times 2^{-72} + (X_1 \times Y_3 + X_2 \times Y_2 + X_3 \times Y_1) \times 2^{-96} + (X_3 \times Y_2) \times 2^{-120} + [(X_1X_2 \times Y_4) \times 2^{-144} + R_1] + [(X_4 \times Y_1Y_2) \times 2^{-144} + R_1] + [(X_2X_3 \times Y_3) \times 2^{-144} + R_1] + R]_{\text{高}n\text{位}} \quad (6)$$

其中，Z 是真值 Z^* 的近似值；$z_f = x_f / y_f$；$zj = xj + yj$。$[X_1X_2 \times Y_4) \times 2^{-144} + R_1]$，$[(X_4 \times Y_1Y_2) \times 2^{-144} + R_1]$，$[(X_2X_3 \times Y_3) \times 2^{-144} + R_1]$是指用硬指令“舍入浮点积”所得的结果，它们分别是用组合金字塔法求$(X_1X_2 \times Y_4) \times 2^{-144} + R_1$，$(X_4 \times Y_1Y_2) \times 2^{-144} + R_1$ 与 $(X_2X_3 \times Y_3) \times 2^{-144} + R_1$ 的结果的高 48 位。这里 $R_1 = 2^{-121}$. 当定点数 $Z_1Z_2Z_3Z_4$ 不须左规时，$R = 2^{-97}$，$n = 96$；当须左规时，$R = 2^{-98}$，$n = 97$。

2 计算步骤

（1）计算数符及阶码部分用硬指令“浮点积”来实现，即先计算

$$(x_f \cdot 2^{xj} \times (1 - 2^{-48})) \times (y_f \cdot 2^{yj} \times (1 - 2^{-48})) \approx z_f \cdot 2^{zj} \times (1 - 2^{-48})$$

然后保留数符和阶码部分。

（2）计算尾数。

①计算尾数时，其组合乘积及对位情况如图 2 所示。这里，所需的单字长乘法和加法的个数分别为 10 个和 9 个，相应地比文[1]的算法分别减少 3 个乘法和 3 个加法。在截取结果时，在最后一位考虑舍入，采用错位连接的技术保留前 98 位。当定点数 $z_1z_2z_3z_4$ 不须左规时，在第 97 位加上舍入 $R = 2^{-97}$。当定点数 $z_1z_2z_3z_4$ 须左规 1 位时，在第 98 位加舍入量 $R = 2^{-98}$，然后左规。

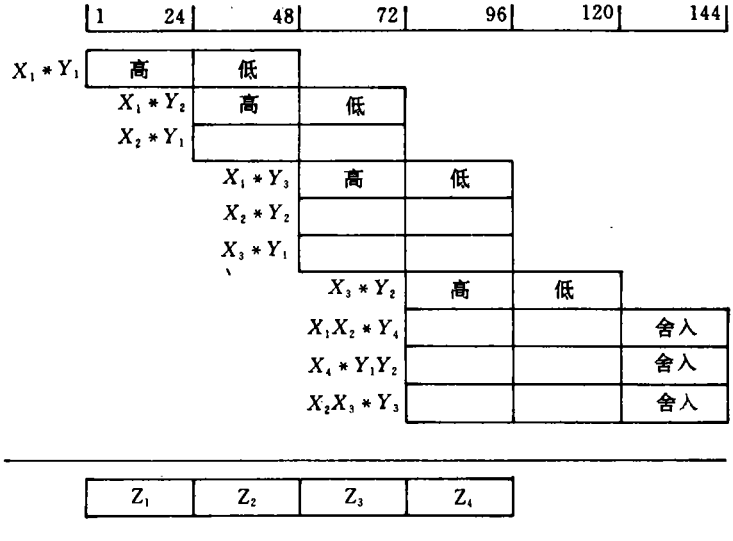

图 2

此处，定点数 $z_1z_2z_3z_4 \leqslant (1-2^{-96})^2+2^{-97}<1$，因此不可能引起右规。

②当定点数 $z_1z_2z_3z_4$ 第 1 位非 0 时，不须左规，由（1）求得的阶码不变。否则，须左规，并将由（1）求得的阶码减 1。

3 误差估计

由公式(6)可知：

(1)Z 中少算的$(X_3\times Y_4)\times 2^{-168}+(X_4\times Y_3Y_4)\times 2^{-192}$是小于 $2\times 2^{-120}=2^{-119}=\triangle_1$；

(2)用硬指令浮点积计算$[(X_1X_2\times Y_4)\times 2^{-144}+R_1]$、$[(X_4\times Y_1Y_2)\times 2^{-144}+R_1]$和$[(X_2X_3\times Y_3)\times 2^{-144}+R_1]$所产生的误差绝对值不超过 $3\times 2^{-121}=\triangle_2$；

(3)定点数 $Z_1Z_2Z_3Z_4$ 不须左规时，式(6)中 $R=2^{-97}$，$n=96$。R 起四舍五入的作用。截取高 96 位的截断误差绝对值不超过 $2^{-97}=\triangle_3$。

综合(1)、(2)、(3)知：当 $Z_1Z_2Z_3Z_4$ 不须左规时，尾数绝对误差

$$\triangle<\triangle_1+\triangle_2+\triangle_3<2^{-97}+2^{-118}<2^{-96} \qquad (7)$$

(4) 定点数 $Z_1Z_2Z_3Z_4$ 须左规时，式 (6) 中 $R=2^{-98}$，$n=97$。R 起四舍五入的作用。截取高 97 位的截断误差绝对值不超过 $2^{-98}=\triangle_4$。综合 (1)、(2)、(4) 知：

当 $Z_1Z_2Z_3Z_4$ 须左规而未左规时，尾数绝对误差

$$\triangle<\triangle_1+\triangle_2+\triangle_4<2^{-98}+2^{-118}<2^{-97}$$

左规 1 位后，尾数的绝对误差

$$\triangle<2^{-97}+2^{-117}<2^{-96} \qquad (8)$$

由(7)、(8)式知，计算结果的尾数绝对误差$\triangle<2^{-96}$，即至少保证有 95 位有效数字。实际上，由于式 (7) 与式 (8) 中的 2^{-118}与 2^{-117}相对 2^{-97}而言非常小，一般均有$\triangle\leqslant 2^{-97}$，即计算结果有 96 位有效数字，也即双字长尾数的所有 96 位均与真值相同。这与实际考核结果是一致的。此算法的精度相对文 [1] 的$\triangle<2^{-95}$，提高了 2 位。

参 考 文 献

1 余龙生．785 机双倍字长浮点算术运算的算法．计算机工程与科学，1981，(1)

2 史应光，李家楷．计算机上高精度四则运算的一种方法．计算数学，1979，1 (4)

3 Clenshaw C W, Olver F J. Beyond Floating Point. Journal of Association for Computing Machinery. April 1984, 1 (31)

（责任编辑　潘　生）