СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

11 - 9680

А.И.Салтыков

ВЫЧИСЛЕНИЕ ЭЛЕМЕНТАРНЫХ ФУНКЦИЙ С ДВОЙНОЙ ТОЧНОСТЬЮ НА БЭСМ-6

1976

11 - 9680

А.И.Салтыков

ВЫЧИСЛЕНИЕ ЭЛЕМЕНТАРНЫХ ФУНКЦИЙ С ДВОЙНОЙ ТОЧНОСТЬЮ НА БЭСМ-6

Введение

В состав постоянной библиотеки мониторной системы "Дубна" на БЭСМ-6 входят программы вычисления элементарных функций с двойной точностью. Из этих программ наиболее часто используются подпрограммы-функции DSQRT(X), DEXP(X), DLOG(X), DSIN(X), DCOS(X), DATAN(X), вычисляющие значения элементарных функций $\sqrt{x}$, e^x, ℓnx, $\sin x$, $\cos x$ и $\mathrm{arctg}x$ с двойной точностью.

Указанные программы составлены на ФОРТРАНе /DSQRT(X) содержит автокодную часть/ и используют арифметические операции над числами с двойной точностью, реализованные в автокодной программе DUBLPREC. Такой способ является неоптимальным как в смысле затрачиваемого счетного времени, так и в смысле объема оперативной памяти, занимаемой программами.

Кроме того, было установлено, что вычисление функций e^x и ℓnx производится всего с 15 верными десятичными знаками, в то время как остальные функции вычисляются с 21 верным знаком /DSQRT(X) дает 23 верных знака/.

Особенно неоптимально реализовано вычисление $\sqrt{x}$, требующее 10000 *мкс* счетного времени, что в несколько раз превышает время счета аналогичной функции на машине М-20 /или БЭСМ-4/, уступающей БЭСМ-6 по быстродействию в десятки раз.

Взамен существующих программ, реализующих вычисление рассмотренных элементарных функций с двойной точностью, предлагаются новые программы, обладающие значительной большей скоростью работы и обеспечивающие большую точность.

Кроме вышеназванных 6 элементарных функций реализована также программа вычисления функции arcsin x с двойной точностью под названием DASIN(X).* Эта подпрограмма-функция не входит в состав стандартных функций с двойной точностью и требует описания в вызывающей программе оператором DOUBLE PRECISION. Время счета функции arcsinx с двойной точностью-1200 *мкс* точность - 21 десятичный знак.

В приведенной ниже таблице указано счетное время для старых и новых программ, а также время счета для аналогичных элементарных функций на машине М-20. Счетное время на БЭСМ-6 измерялось с помощью системной программы CTIME /см. /1/, , стр. 237/, данные для М-20 взяты из /2/.

При обращении к вычислению функции от "плохого" аргумента теперь выдается типовая диагностика: ** BAD ARGUMENT AT <наименование функции> с выходом на запрещенную команду. При этом индекс регистр 13 указывает адрес возврата в вызывающую программу.

Таблица времени счёта

Функция		$\sqrt{x}$	e^x	$\ell n x$	sin x	cos x	arctg x
Счётное время в мкс	старые	10000	9000	12000	14000	14000	6000
	новые	150	700	800	800	800	900
	М-20	4000	41000	393000	36000	40000	55000

Как видно из таблицы, счетное время удалось сократить на порядок, а для функции $\sqrt{x}$ - более чем в 50 раз.

Новые программы обеспечивают точность не менее 21 верного десятичного знака. Функция $\sqrt{x}$ вычисляется с 23 верными знаками. Место в оперативной памяти, занимаемое программами, сократилось в 2-3 раза.

1. *Реализация арифметических операций над числами с двойной разрядностью*

С целью сокращения счетного времени мы отказались от использования программы DUBLPREC, реализующей выполнение арифметических операций над числами с двойной точностью.

Взамен была составлена специальная программа DPARITHM, реализующая арифметические операции над числами с двойной разрядностью. Эта программа позволяет также вычислять значения многочлена и производить ряд вспомогательных операций.

Программа DPARITHM оперирует с числами, имеющими мантиссу, состоящую из 80 двоичных разрядов, и лежащими в диапазоне порядков от -80 до +47. Для того, чтобы избежать получения машинных нулей, когда результат меньше 2^{-65} по модулю, все числа предварительно умножены на масштабный коэффициент 2^{16}. Результат операции также получается с маштабным коэффициентом 2^{16}. Такие числа мы будем в дальнейшем называть промежуточными.

2. *Общая схема вычисления функций с двойной точностью*

Вычисление функций с двойной точностью, кроме DSQRT(X), производится по единообразной схеме с использованием программы DPARITHM, которая оперирует с промежуточными числами. Опишем основные этапы процесса вычисления рассматриваемых функций.

а/ Преобразование аргумента из формата числа с двойной точностью в промежуточное число.

Сначала производится отделение порядка числа с двойной точностью от мантиссы. Затем производится приведение аргумента к "рабочему" диапазону /своему для каждой функции/. В процессе приведения анализируются случаи, когда функцию надо вычислять особо /например, при $|X| < 2^{-39}$ полагается DSIN(X) = X /.

Здесь же производится выход на диагностику в случаях, когда аргумент лежит вне области определения функции или значение функции выходит за диапазон допустимых чисел.

б/ Вычисление значения функции от промежуточного аргумента.

После приведения аргумента к рабочему диапазону и представления его в виде промежуточного числа производится вычисление аппроксимирующего выражения. В качестве аппроксимирующего выражения выбран полином для DSIN(X) и DCOS(X) и дробно-рациональное выражение для остальных функций. Коэффициенты аппроксимирующих выражений были взяты из таблиц /3/, причем каждый раз выбиралось выражение с наименьшим числом коэффициентов, обеспечивающее относительную погрешность вычисления /без учета машинной погрешности/ не более 10^{-24}.

Характерной особенностью используемых аппроксимирующих выражений является то, что все они содержат многочлены от квадрата аргумента, т.е. выражения вида $P(X^2)$, где P - многочлен.

в/ Приведение результата к форме числа с двойной точностью.

После вычисления e^x от промежуточного аргумента производится корректировка порядка и приведение результата к форме числа с двойной точностью. Для остальных функций производится преобразование результата из формы промежуточного числа в форму числа с двойной точностью.

Все рассматриваемые подпрограммы-функции "портят" состояние индекс-регистров 8÷14.

3. Вычисление DSQRT(X)

Пусть $X = 2^{64P} \cdot x$ - число с двойной точностью на БЭСМ-6. Здесь $-63 \le P \le 63$ - старшие разряды порядка, x - число с 80-разрядной мантиссой: $x = x_1 + x_2 \cdot 2^{-40}$, лежащее в обычном диапазоне чисел БЭСМ-6. Тогда $\sqrt{X} = 2^{32P} \cdot \sqrt{x}$. Вычисление $\sqrt{x}$ производится по алгоритму, описанному в /2/ /стр. 96/. Этот алгоритм эквивалентен итерации Герона

$$y_{n+1} = \frac{1}{2}\left(y_n + \frac{x}{y_n}\right),$$

где $y_n = \sqrt{x_1}$ и вычисление производится с учетом 80-разрядной мантиссы числа x.

Для вычисления $\sqrt{x_1}$ используется экстракод *50. При этом x_1 предварительно приводится к диапазону порядков от -25 до 39.

В случае $x < 0$, что соответствует $X < 0$, происходит выход на диагностику и затем на стоп по запрещенной команде /запр.ком./.

После вычисления $\sqrt{x}$ производится приведение результата к виду числа с двойной точностью.

Значение корня получается с 23 верными знаками. Чаще всего значение корня получается с недостатком. Если $x = y^2$ и $x < 2^{40}$, то корень вычисляется точно.

4. Вычисление DEXP(X)

Вычисление e^X, где X - число с двойной точностью, производится по формуле

$$e^X = 2^{X \cdot \frac{1}{\ln 2}} = 2^M \cdot (2^z)^2,$$

где

$$M = [X \cdot \frac{1}{\ln 2}], \quad z = \frac{1}{2}\{X \cdot \frac{1}{\ln 2}\}.$$

Для вычисления 2^z используется дробно-рациональное приближение

$$2^z \approx \frac{Q(z^2) + zP(z^2)}{Q(z^2) - zP(z^2)},$$

где P и Q - многочлены степени 3 с коэффициентами, взятыми из /3/ /таблица 1325/.

В случае, когда порядок результата превышает 4095, происходит выход на диагностику. Если порядок результата меньше -4096, то в качестве результата выдается машинный ноль.

5. Вычисление DLOG(X)

Пусть $X = 2^N \cdot x$ - число с двойной точностью, где $-4096 \le N \le 4095$ - порядок, $1/2 \le x < 1$ - мантисса. Тогда

$\ell n X = N\ell n 2 + \ell n x$.

Вычисление $\ell n x$ производится по формуле

$$\ell n x = -\frac{\ell n 2}{2} + z \cdot \frac{P(z^2)}{Q(z^2)},$$

где P и Q - многочлены степени 4 с коэффициентами, взятыми из /3/ /таблица 2707/ и $z = \frac{x-\sqrt{2}/2}{x+\sqrt{2}/2}$.

Результат получается с 21 верным десятичным знаком. Исключение составляют значения X вблизи 1, когда точность резко падает из-за увеличения относительной погрешности величины $X-1$ /вблизи $X = 1$ $\ell n X \approx X-1$/.

При $X \le 0$ происходит выход на диагностику и стоп по запрещенной команде.

6. Вычисление DSIN(X) и DCOS(X).

При вычислении sinx с двойной точностью аргумент приводится к диапазону $0 \le x \le \frac{\pi}{2}$, после чего производится вычисление по формуле

$$\sin\frac{\pi}{2}z \approx zP(z^2).$$

Здесь $0 \le z \le 1$ и P - многочлен 10-й степени с коэффициентами из таблицы 3343 в /3/.

Практически значение sinx получается с 21 верным знаком. Исключение составляют большие по модулю значения x /превосходящие приблизительно 1000/, когда после отбрасывания периодов у аргумента остается мало верных знаков.

В случае $|x| > 2^{80}$ после отбрасывания периодов у аргумента не остается ни одного верного знака. В этом случае программа выходит на диагностику.

При $|x| < 2^{-39}$ в качестве sinx выдается x.

Вычисление cosx с двойной точностью сводится к вычислению $\sin(\frac{\pi}{2} - x)$. При этом сначала у аргумента отбрасываются периоды и только затем производится

переход от x к $\frac{\pi}{2} - x$. Это обеспечивает выполнение равенства

$$\sin^2 x + \cos^2 x = 1$$

с точностью до 21 десятичного знака при любых значениях аргумента.

В случае $|x| > 2^{80}$ программа выходит на диагностику.

7. Вычисление DATAN(X) и DASIN(X).

Вычисление arctgx с двойной точностью производится по формуле

$$t = \operatorname{arctg} x = \begin{cases} \operatorname{arctg} z, & \text{где } z = x, \text{ при } x \le 2-\sqrt{3} \\ \operatorname{arctg} z + \frac{\pi}{6}, & \text{где } z = \dfrac{x - \frac{1}{\sqrt{3}}}{1 + x\frac{1}{\sqrt{3}}}, \text{ при } 2-\sqrt{3} < x \le 1. \end{cases}$$

При этом x приводится к интервалу $0 \le x \le 1$ путем смены знака при $x < 0$ и перехода от x к $\frac{1}{x}$ при $|x| > 1$.

Вычисление arctgz производится по формуле

$$\operatorname{arctg} z \approx z \cdot \frac{P(z^2)}{Q(z^2)},$$

где P и Q - многочлены 5-й степени с коэффициентами из /3/ /таблица 5058/.

Для получения окончательного результата надо произвести вычисление $\frac{\pi}{2} - t$ при $|x| > 1$ и учесть знак аргумента.

При $|x| \le 2^{-39}$ выдается $\operatorname{arctg} x = x$, а при $|x| > 2^{79}$ полагается $\operatorname{arctg} x = \frac{\pi}{2} \cdot \operatorname{sgn} x$.

Вычисление arcsinx производится по формуле

$$\arcsin x = \operatorname{arctg} \frac{x}{\sqrt{1 - x^2}}, \qquad |x| < 1.$$

При $|x| = 1$ выдается $\arcsin x = \frac{\pi}{2} \operatorname{sgn} x$. Если $|x| \le 2^{-39}$, то полагается $\arcsin x = x$.

В случае $|x| > 1$ производится выдача диагностики.

Заключение

Методика вычисления элементарных функций с двойной точностью, рассмотренная в данной работе, позволила в среднем на порядок сократить время счета по сравнению с прежними вариантами программ. Отметим основные факторы, позволившие получить указанный эффект.

Во-первых, новые программы написаны на автокоде, а не на ФОРТРАНе. Этот фактор наиболее существенно сказался при вычислении тех функций, которые требуют выполнения нестандартных операций над числами с двойной точностью. Например, при вычислении DSQRT(X), DEXP(X) и DLOG(X) необходимо выделять порядок числа с двойной точностью, что трудно сделать средствами ФОРТРАНа. Поэтому в старых вариантах соответствующих программ их автору пришлось идти "окольными путями", что значительно увеличило счетное время. При вычислении DATAN(X) этот фактор сказался не столь существенно, поскольку алгоритм вычисления этой функции не требует выполнения нестандартных операций.

Во-вторых, в новых программах использованы более эффективные аппроксимирующие выражения, взятые из /3/.

В-третьих, в новых вариантах программ вместо арифметических операций над числами с двойной точностью используются операции над числами с 80-разрядной мантиссой, лежащими в обычном диапазоне чисел БЭСМ-6.

Отметим, что данная методика вычисления может применяться и для других функций с двойной точностью на БЭСМ-6 / например, гамма-функция, функции Бесселя и др./.

Автор благодарит Г.Л.Мазного за консультации и дружескую поддержку.

Литература

1. *Г.Л.Мазный. Мониторная система "Дубна", ОИЯИ, 11-5974, Дубна, 1971.*
2. *Библиотека стандартных программ. Ред. М.Шура-Бура, ЦБТИ, М., 1961.*
3. *J.F.Hart a.o. Computer Approximations, N.Y., John Wiley and Sons, 1968.*

Рукопись поступила в издательский отдел
2 апреля 1976 года.