ПОВЫШЕНИЕ ТОЧНОСТИ ЧИСЛЕННОГО ИНТЕГРИРОВАНИЯ ГРАВИТАЦИОННОЙ СИСТЕМЫ N ТЕЛ С ПОМОЩЬЮ СДВОЕННЫХ ЧИСЕЛ ДВОЙНОЙ ТОЧНОСТИ

М.О. Субботин, А.В. Кодуков

Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»

Аннотация. С начала эры компьютеризации небесной механики до конца XX века чисел двойной точности хватало для расчёта орбит небесных тел. По мере появления более точных наблюдений и более детализированных моделей появилась необходимость и в более точной арифметике. Ранее для этого использовались процессоры с числами четверной точности, но на данный момент поддерживаются только числа расширенной точности. Использование этих чисел представляется неудобным из-за несовместимости с некоторыми архитектурами и языками программирования. В данной работе исследована возможность замены чисел расширенной точности арифметикой double-double с некоторыми оптимизациями, чтобы сохранить быстродействие и получить нужную точность.

Ключевые слова: double-double, машинная арифметика, задача N тел, методы численного интегрирования

Задача N тел

В данной работе решается гравитационная задача N тел (материальных точек) в применении к Солнечной системе. Задача определяется следующим образом: в пространстве находятся N материальных точек с известными массами, положениями и скоростями на начальный момент времени. Попарное взаимодействие точек подчиняется Закону всемирного тяготения. Требуется найти положения материальных точек в последующие моменты времени.

Рассмотрение системы из одного и двух тел не представляют интереса, так как такая система полностью описывается законами Ньютона. Для $N \geq 3$ не существует аналитических решений в общем виде. Существует аналитическое решение для трёх тел в виде рядов [1]. Эти ряды сходятся для любого момента времени, с любыми начальными условиями, но сходятся они крайне медленно и поэтому этот подход не применим на практике. Следовательно, для $N \geq 3$ решения необходимо находить численными методами. Важной задачей является сведение ошибки численного метода к минимуму.

Взаимодействие тел описывается Законом всемирного тяготения. Суммарное ускорение тела, с помощью которого вычисляется перемещение, можно рассчитать так:

$$\vec{a}_n = \frac{\vec{F}_n}{m_n} = -G \sum_{k \neq n} m_k \frac{\vec{r}_n - \vec{r}_k}{|\vec{r}_n - \vec{r}_k|^3}$$

Для проверки корректности численного интегрирования обычно применяют два метода: сравнение с аналитическим решением и вычисление инвариантов. Сравнение с аналитическим решением не представляется возможным, т.к. для $N \geq 3$ его не существует. Но в системе N тел есть три величины, которые не меняются со временем.

Инвариантами данной системы являются полная энергия системы (Закон сохранения энергии), сумма моментов импульса (Закон сохранения момента импульса) и барицентр системы (если на механическую систему не действуют внешние силы, то её центр масс движется с постоянной по величине и направлению скоростью).

Методы численного интегрирования

Уравнение вида ($\vec{r}'' = \vec{a}$) необходимо свести к виду, соответствующему задаче Коши. В задаче Коши считаются известными начальное состояние системы и функция расчёт производной состояния.

Для сведения уравнения

$$\frac{\partial^2 \vec{r}_n}{\partial t^2} = a_n = -G \sum_{k \neq n} m_k \frac{(\vec{r}_n - \vec{r}_k)}{|\vec{r}_n - \vec{r}_k|^3}$$

к системе первого порядка вводится величина скорости тела, с помощью которой получается следующая система:

$$\begin{cases} \dfrac{\partial \vec{v}_n}{\partial t} = a_n \\ \dfrac{\partial \vec{r}_n}{\partial t} = \vec{v}_n \end{cases}$$

Методы Рунге-Кутты – большой класс численных методов решения задачи Коши для обыкновенных дифференциальных уравнений. В данной задаче для сравнения были выбраны метод Рунге-Кутты 4 порядка и метод Дормана-Принса 8 порядка [2].

Машинная арифметика

При вычислении скорости тела выполняются в большом количестве операции сложения, умножения, деления, взятие корня. Из-за чего возникает потребность в использовании представления чисел с плавающей точкой с высокой точностью. На рис 1. перечислены представления чисел с плавающей точкой, над которыми проводилось исследование.

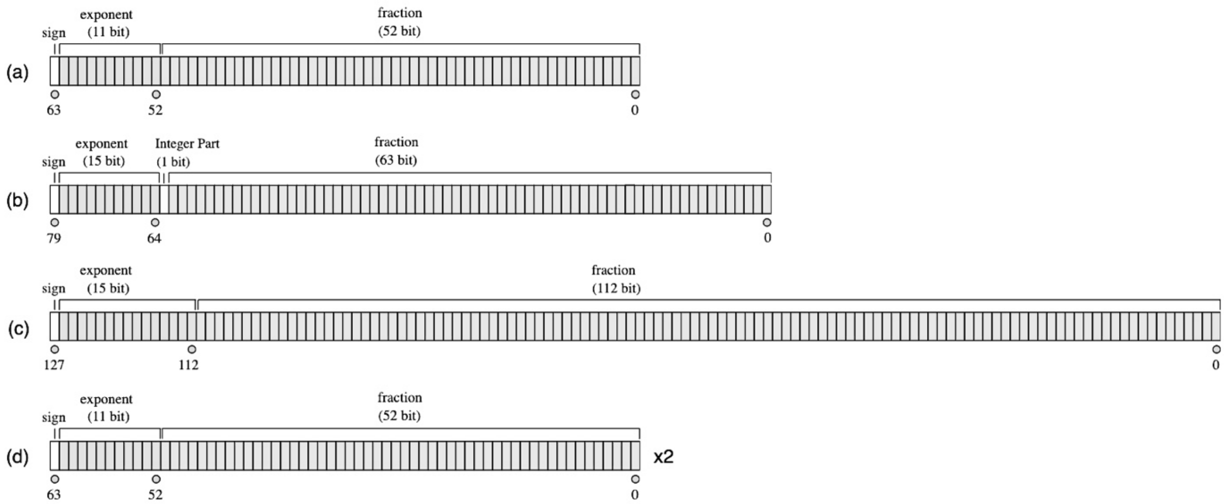

Рис 1. Битовые представления чисел с плавающей точкой различной точности: a) число двойной точности (double) [3], b) число расширенной точности (Extended/long double) [4], c) число четверной точности (quadruple) [5], d) сдвоенное число двойной точности (double-double) [6].

Double имеет недостаточную точность (количество битов в мантиссе) для этой задачи. Точности long double хватает, но не во всех языках программирования и архитектурах процессоров есть поддержка этого типа. Quadruple имеет избыточную точность, он реализован программным путем и из-за этого операции над этим типом выполняются медленно.

В свою очередь double-double, как и quadruple имеет программную реализацию [7] (т.е. не зависит от архитектуры и языка программирования) и выражается как сумма двух компонент double (большого и маленького числа) $x = x_h + x_l$, где $|x_l| \leq \frac{1}{2} ulp(x_h)$. Double-double имеет меньшую точность, чем quadruple, поэтому ожидаемо операции над ним должны работать быстрее.

Особый интерес представляет сравнение точности и скорости типов long double и double-double в данной задаче.

Сравнение результатов численного интегрирования

Для исследования была выбрана система, состоящая из 8 тел (планета и 7 спутников, рис. 2). Численное интегрирование орбит осуществляется за 10^5 шагов.

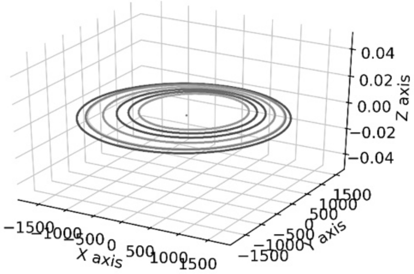

Рис 2. Траектория передвижения тел. Вокруг планеты вращаются 7 спутников.

Интегрирование с double-double оказалось медленнее (таблица 1) интегрирования с long double в 3.67 и 3.60 раз для методов Рунге-Кутты и Дормана-Принса соответственно. Также методы с double-double оказались быстрее методов с quadruple в 3.04 и 2.4 раза.

Таблица 1

Сравнение времени работы методов численного интегрирования

	RK4 (c)	DOPRI8 (c)
Double	0.7542	3.275
Long double	1.266	5.925
Quadruple	14.14	51.33
Double-double	4.653	21.32

Для определения ошибки методов интегрирования вычисляется относительная ошибка изменения величин (инвариантов), которые указаны в таблице 2.

Таблица 2

Порядок накопленной относительной ошибки инвариантов

	RK4			DOPRI8		
	Энергия $\frac{\lvert E - E_0 \rvert}{E_0}$	Момент импульса $\frac{\lvert L - L_0 \rvert}{L_0}$	Барицентр $\lvert X - X_0 \rvert$	Энергия $\frac{\lvert E - E_0 \rvert}{E_0}$	Момент импульса $\frac{\lvert L - L_0 \rvert}{L_0}$	Барицентр $\lvert X - X_0 \rvert$
Double	$\sim 10^{-14}$	$\sim 10^{-14}$	$\sim 10^{-23}$	$\sim 10^{-14}$	$\sim 10^{-14}$	$\sim 10^{-23}$
Long double	$\sim 10^{-16}$	$\sim 10^{-18}$	$\sim 10^{-26}$	$\sim 10^{-18}$	$\sim 10^{-18}$	$\sim 10^{-26}$
Quadruple	$\sim 10^{-17}$	$\sim 10^{-18}$	$\sim 10^{-41}$	$\sim 10^{-32}$	$\sim 10^{-32}$	$\sim 10^{-42}$
Double-double	$\sim 10^{-16}$	$\sim 10^{-18}$	$\sim 10^{-39}$	$\sim 10^{-30}$	$\sim 10^{-31}$	$\sim 10^{-39}$

Инварианты с double-double имеют порядок относительной ошибки сравнимый с порядком инвариантов quadruple (отличаются на 1-3 порядка). В интегрировании с методом Дормана-Принса quadruple и double-double значительно точнее, чем long double на 11-16 порядков. В методе Рунге-Кутты получается одинаковая ошибка в энергии и моменте импульса для всех типов из-за того, что сам метод имеет меньшую точность, чем возможная точность ошибки в типах.

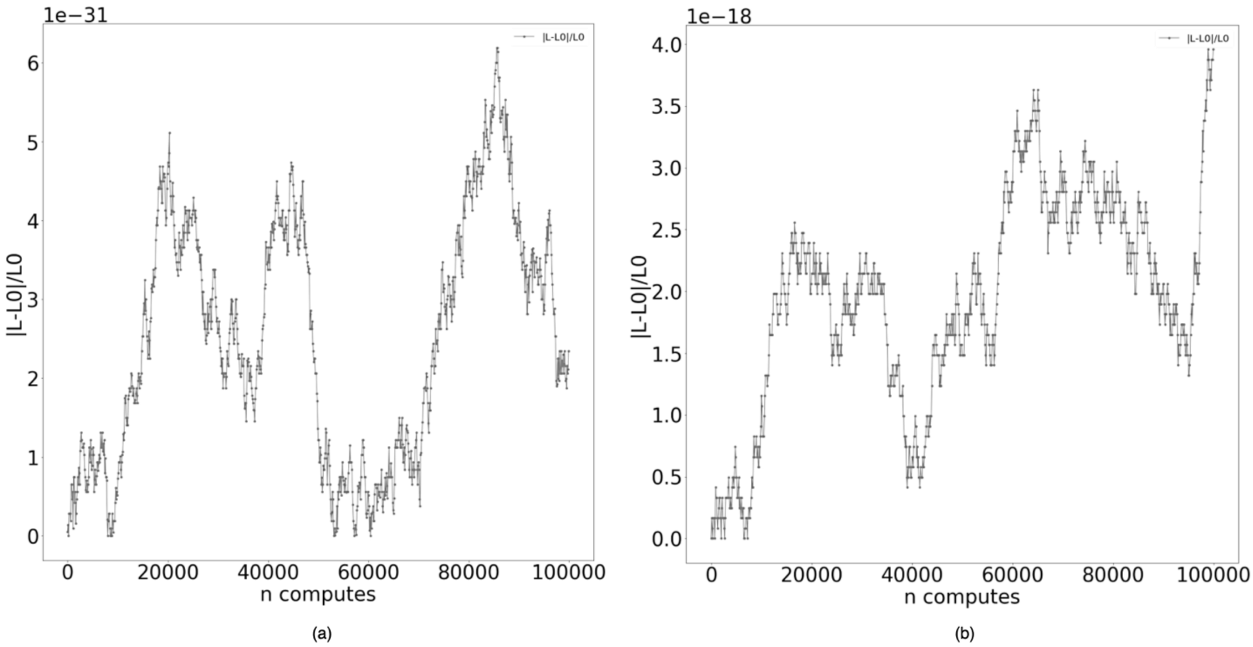

Рис 3. Графики относительного изменения суммы моментов импульса для a) double-double, b) long double в интегрировании с помощью метода Дормана-Принса.

На рис 3. Относительное изменение инвариантов системы не монотонно. Это объясняется тем, что вычисления инвариантов такой системы имеют хаотичный характер. Главным показателем является максимальная накопленная относительная ошибка.

Вывод

Тип double-double оказался применим для данной задачи. Интегрирование на double-double показывает сравнимую точность с quadruple. Интегрирование с использованием double-double быстрее чем с использованием quadruple примерно в 3 раза, но медленнее чем с long double в 3.5 раза. Дальнейшие исследования будут направлены на приближение скорости численного интегрирования с double-double к скорости с long double. Этого можно добиться путем выполнения некоторых вычислений на double. Также планируется расширение транслятора предметно-ориентированного языка Landau [8] арифметикой double-double.

Список литературы

1. Karl F. Sundman. Mémoire sur le problème des trois corps. Acta Math. 36 105 - 179, 1913.
2. https://gitlab.iaaras.ru/iaaras/abmd/blob/master/src/rk.c#L130 (дата обращения: 14.04.2021)
3. https://en.wikipedia.org/wiki/Double-precision_floating-point_format (дата обращения: 14.04.2021)
4. https://en.wikipedia.org/wiki/Extended_precision (дата обращения: 14.04.2021)
5. https://en.wikipedia.org/wiki/Quadruple-precision_floating-point_format (дата обращения: 14.04.2021)
6. Jean-Michel Muller, Nicolas Brisebarre, Florent de Dinechin. Handbook of Floating-Point Arithmetic 2010th Edition
7. Hida, Yozo & Li, Sherry & Bailey, David. (2008). Library for Double-Double and Quad-Double Arithmetic.
8. I. Dolgakov, D. Pavlov. "Landau: language for dynamical systems with automatic differentiation". Zap. Nauch. Sem. POMI 485 (2019)