Московский Физико-Технический Институт

Лаба по флоатам

Выполнил:
Студент 1 курса ФАКТ
Группа Б03-504
Подмосковнов Лев

Долгопрудный
2025

0. unsigned int -> binary

```
void ui_to_bin(unsigned int a) {
    for (int i = 32 - 1; i >= 0; i--) {
        cout << ((a >> i) & 1);
        if (i != 0 && i % 8 == 0) {
            cout << " ";
        }
    }
}
```

Функция для перевода и печати в двоичную систему переменную типа unsigned int.

1. float -> binary

```
union fu {
 float f;
 unsigned int u;
};

void ui_to_bin(unsigned int a) {
    for (int i = 32 - 1; i >= 0; i--) {
        cout << ((a >> i) & 1);
        if (i == 31 || i == 23) {
            cout << "|";
        }
        if (i != 0 && i % 4 == 0) {
            cout << " ";
        }
    }
}
```

В структуру fu записываем в f наше число. Теперь нужно посмотреть на биты float через unsigned int.

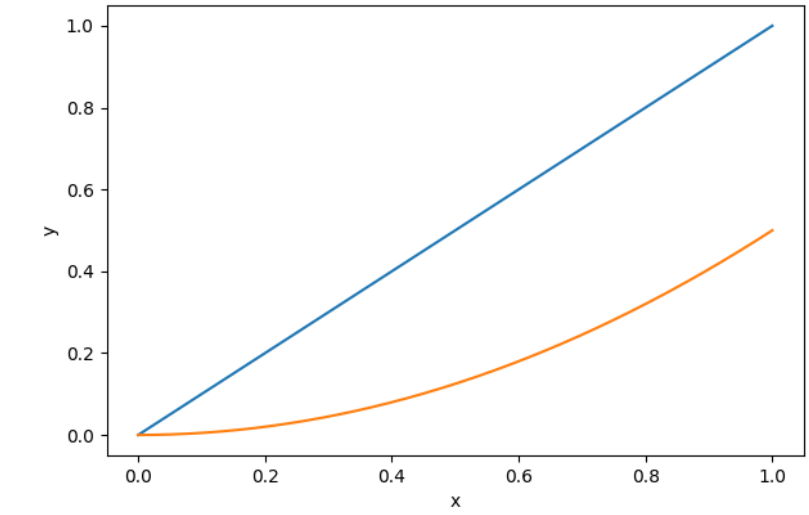
2. Переполнение мантисы

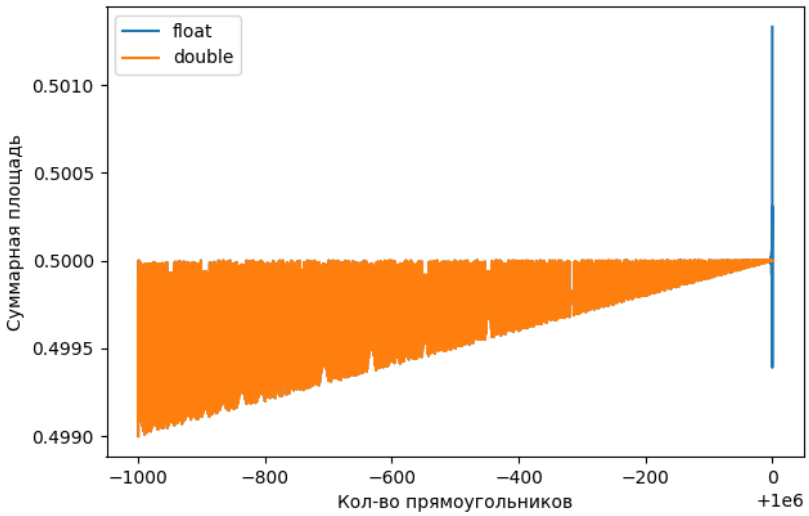
В ходе эксперимента найдено, что последняя степень 10, которая не переполняет мантису 11. Когда хотим положить в float 2^{12} получается число 99999997952.00.

3. Бесконечный цикл

В ходе эксперимента было замечено, что к число 16777216.0 в float нельзя добавить единицу.

4. Численное интегрирование





При увелечении количества столбцов результат становится точнее и в float, и в double. Float ломается быстрее чем double.

5*. Антипереполнение

Subnormal / denormal — это числа с очень маленьким модулем, которые находятся между нулём и самым маленьким нормальным числом IEEE-754. Они имеют нулевой (biased) экспонент-поле и ненулевую мантиссу; обеспечивают плавное (gradual) антипереполнение. Без них некоторые разности могли бы сразу дать 0.

Underflow (антипереполнение) — ситуация, когда результат арифметики меньше минимального нормального значения. С появлением subnormals результат может быть subnormal (потеря точности), а без них — прямо 0.

DAZ (Denormals-Are-Zero) — если входной операнд (операнты) является денормалом, он заменяется на ноль до выполнения операции (то есть входные денормалы “приравниваются к 0”).

FTZ (Flush-To-Zero) — если результат операции получился денормалом, вместо него возвращается 0.

При DAZ/FTZ OFF вы часто увидите, что операция с денормалами значительно медленнее