# Лекция №1
# по курсу
# «Машинная арифметика в рациональных чисел»

Москва, 2020

# Литература


1. Overton, Michael L. Numerical computing with IEEE floating point arithmetic
2. Behrooz Parhami. Computer arithmetic
3. Koren Izrael. Computer arithmetic algorithms/ 2nd ed.
4. Поспелов Д.А. Арифметические основы вычислительных машин дискретного действия. Учеб. пособие. - М.: Изд-во "Высш. школа", 1970. - 308 с.
5. Яглом И., Системы счисления. Журнал Квант

# Компьютерная арифметика

Было бы ошибкой считать, что компьютерная арифметика необходима только разработчикам процессоров. Мы рассмотрим дальше примеры, как более эффективно точнее составлять расчётные программы, избегать вычислительных ошибок, свойственных арифметики с плавающей точкой.

Основные вопросы предмета компьютерная арифметика – это:

1. Разработка эффективных цифровых схем.
2. Ускорение арифметических операций и вычисление специальных функций.
3. Разработка алгоритмов быстрого выполнения арифметических операций.
4. Анализ ошибок округления,
5. Аппаратная реализация.
6. Тестирование, верификация программ

# Требования к системам счисления

1. Возможность представления чисел в заданном диапазоне
2. Однозначность представления
3. Простоту записи
4. Удобство работы человека с машиной
5. Трудоёмкость выполнения арифметических операций
6. Экономичность системы (количество элементов, необходимое для представления многоразрядных чисел)
7. Удобство аппаратной реализации

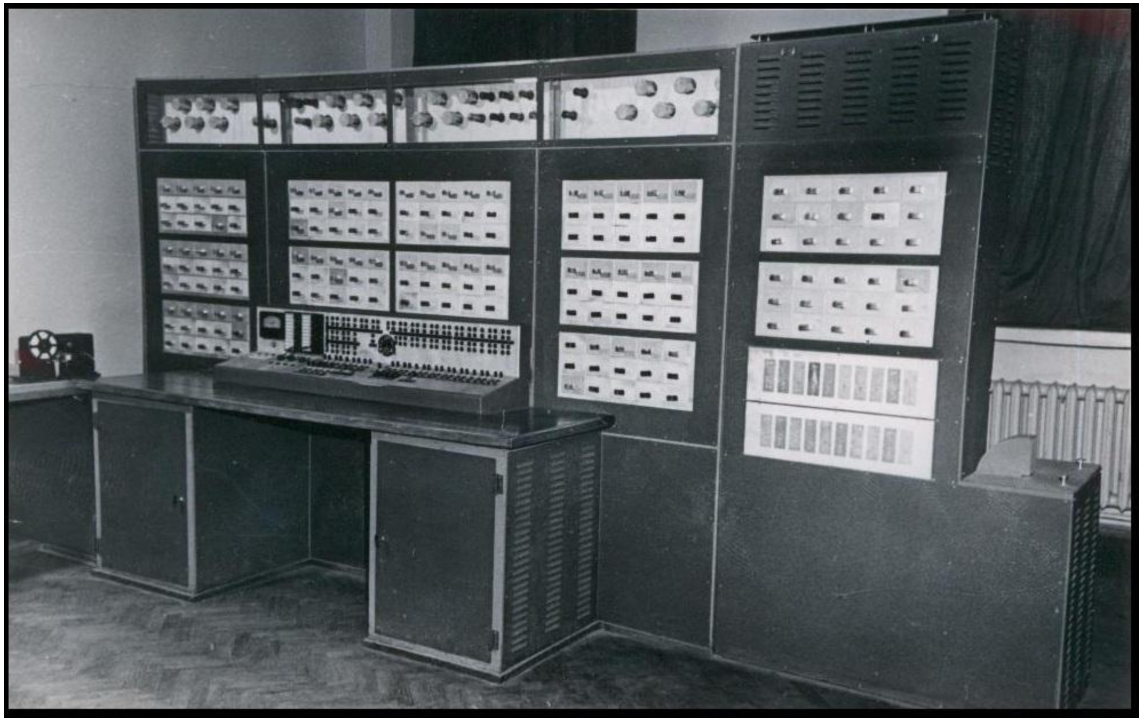
# Вычислительная машина «Сетунь»

Брусенцов Николай Петрович
(1925 г – 2014 г)

60-е года прошлого века
(-1, 0, 1)

-5 = (-1)(0)(-1)

# Сетунь – первый в мире троичный компьютер



# Представление чисел в системах счисления

Любое число в позиционной системе счисления с основанием $q$ можно представить в виде

$$a_{n-1}q^{n-1} + \ldots + a_1 q^1 + a_0,$$

где

$a_0, \ldots a_{n-1}$ – цифры,

$q$ – основание системы счисления

# Представление в смешанной системе счисления

Рассмотрим упорядоченный набор из $n$ целых чисел

$$\rho = [r_1, r_2, \ldots, r_n],$$

компоненты которого $r_1, r_2, \ldots, r_n$ назовем *основаниями*. Пусть $R$ есть произведение оснований, т. е.

$$R = \prod_{i=1}^{n} r_i.$$

Хорошо известно (например, см. [Szabó, Tanaka, 1967, с. 41]), что каждое целое число $s$, такое, что

$$0 \leqslant s < R,$$

можно единственным образом представить в виде

$$s = d_0 + d_1(r_1) + d_2(r_1 r_2) + \ldots + d_{n-1}(r_1 r_2 \ldots r_{n-1}),$$

# Представление в смешанной системе счисления

$$0 \leqslant d_i < r_{i+1}, \quad i = 0,\ 1,\ \ldots,\ n-1.$$

Заметим, что основная роль $r_n$ служить границей для $d_{n-1}$.

*Упорядоченный набор цифр* $d_0, d_1, \ldots, d_{n-1}$ для данного $s$ записывается в виде

$$\langle s \rangle_\rho = \langle d_0,\ d_1,\ \ldots,\ d_{n-1} \rangle.$$

Например, если $\rho = [2, 3, 5]$, то $R = 30$: Следовательно, из

$$29 = 1 + 2\,(2) + 4\,(2 \cdot 3)$$

имеем $d_0 = 1$, $d_1 = 2$, $d_2 = 4$. Отсюда

$$\langle 29 \rangle_\rho = \langle 1,\ 2,\ 4 \rangle.$$

# Представление в смешанной системе счисления

S = d0 + r1(d1 + d2 (r2) + … + dn-1 (r2*r3*…*rn-1) )

t1 = (d1 + d2 (r2) + … + dn-1 (r2*r3*…*rn-1) )

S = d0 + r1*t1

d0 =

S – d0 = r1(d1 + d2 (r2) + … + dn-1 (r2*r3*…*rn-1) )

(S – d0)*r1^(-1) = d1 + d2 (r2) + … + dn-1 (r2*r3*…*rn-1)

# Система счисления с отрицательным основанием

Система счисления с отрицательным основанием $-q$, множество цифр $[0, q-1]$.

$$\sum_{i} x_i \cdot (-q)^i = \sum_{i\ \text{четн}} x_i \cdot (q)^i - \sum_{i\ \text{не четн}} x_i \cdot (q)^i$$

# Экономичность систем счисления

Четкое размещение максимума экономичности может быть установлено методом последующих рассуждений. **Пусть имеется п символов для записи чисел, а основание системы счисления р.** Тогда количество разрядов числа k = п/р, а полное количество чисел (N), которые могут быть составлены, равно:

$$N = p^{\frac{n}{p}}. \qquad (4.10)$$

Если считать N(p) непрерывной функцией, то можно отыскать то значение $p_т$, при котором iN воспринимает наибольшее значение. Функция имеет вид, представленный на рис.4.3.

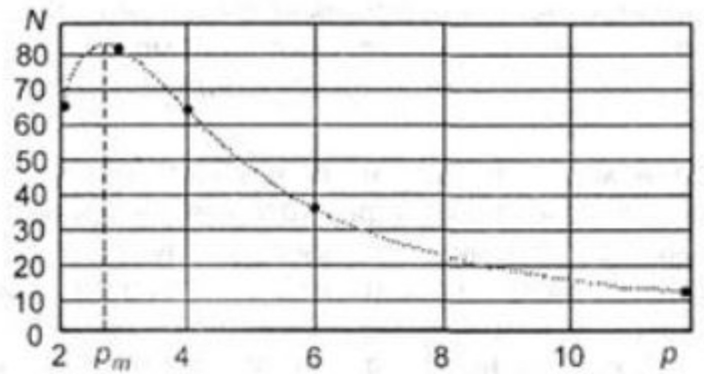



Рис. 4.1. Зависимость количества чисел от основания системы счисления при использовании 12-ти возможных цифр для записи чисел

# ДОПОЛНИТЕЛЬНЫЙ КОД

Дополнительный код позволяет заменить операцию вычитания на операцию сложения и сделать операции сложения и вычитания одинаковыми для знаковых и беззнаковых чисел, чем упрощает архитектуру ЭВМ. В англоязычной литературе обратный код называют первым дополнением, а дополнительный код называют

Положительное целое число $x$, где $0 \le x \le 2^{31} - 1$ записывается как $x$, отрицательное число $-y$, где $1 \le y \le 2^{31}$ как положительное $2^{32} - y$.

При записи числа в дополнительном коде старший разряд является знаковым. Если его значение равно 0, то в остальных разрядах записано положительное двоичное число, совпадающее с прямым кодом.

Двоичное 3-разрядное число со знаком в дополнительном коде может представлять любое целое в диапазоне от $-4$ до $+3$. Если старший разряд равен нулю, то наибольшее целое число, которое может быть записано в оставшихся 2 разрядах, равно $2^2 - 1 = 3$.

# ДОПОЛНИТЕЛЬНЫЙ КОД

| Десятичное представление | Двоичное | Дополнительный код |
|---|---|---|
| 0 | 000 | 0 |
| 1 | 001 | 1 |
| 2 | 010 | 2 |
| 3 | 011 | 3 |
| 4 | 100 | -4 |
| 5 | 101 | -3 |
| 6 | 110 | -2 |
| 7 | 111 | -1 |

Умножение может дать целочисленное переполнение.

Целочисленное деление на ноль обычно приводит к завершению программы и сообщению об ошибке для пользователя

# ДОПОЛНИТЕЛЬНЫЙ КОД

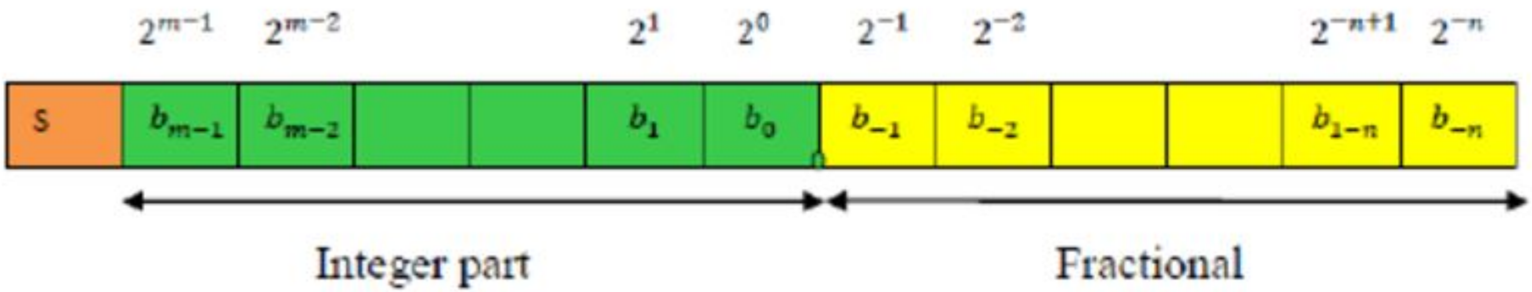
**Упражнение 1.**

Покажите, что если целое число x между и представлено в дополнительном коде, самый левый бит равен 1, если x отрицателен, и 0, если x равен 0 или положительный.

**Упражнение 2.**

Простой способ преобразовать представление неотрицательного целого числа x к представлению в дополнительный код к -x начинается с изменения всех нулевых битов на единичные и всех единичных бит в нулевые. Еще один шаг необходим для завершения процесса. Какой шаг и зачем?

# Формат с фиксированной точкой




11/2 = 0 000000000000101 1000000000000000

$$x = (0000\ 0000.0000\ 1001)_2,$$

$$y = (1001\ 0000.\ 0000\ 0000)_2$$

## 2. ФОРМАТ С ПЛАВАЮЩЕЙ ТОЧКОЙ

В формате представления чисел с плавающей точкой имеем:

$$x = (-1)^s \cdot m \cdot q^e,$$

где

$s \in \{0,1\}$ - знак числа,

$m$ – мантисса, $m \geq 0$,

$e$ – экспонента (целое число).

Число с плавающей запятой имеет четыре компонента: знак $s$, мантиссу $m$, основание системы счисления $q$ и показатель $e$. Вместе эти четыре компонента представляют собой число.

Мантисса числа $x$ имеет $n$ значащих цифр.

Специальный случай, когда $m = 0$ служит для представления нуля.

# Нормализованный формат с плавающей точкой

$$x = \pm S \cdot 10^{E}, \quad 1 \leq S < 10$$

$$x = \pm S \cdot 2^{E}, \quad 1 \leq S < 2$$

$0{,}123 = 0{,}123 \cdot 10^{0}$

$0{,}123 = 123 \cdot 10^{-3}$

$0{,}123 = 1.23 \cdot 10^{-1}$

# Формат с плавающей точкой

Двоичное представление мантиссы имеет вид:

$$S = (b_0 . b_1 b_2 b_3 \ldots)_2, c \; b_0 = 1$$

Например,

$$\frac{13}{2} = 6 + \frac{1}{2} = (1.101) \cdot 2^2$$

Так как бит $b_0 = 1$, то

$$S = (1 . b_1 b_2 b_3 \ldots)_2,$$

Часть числа, следующая после двоичной точки называют дробной частью мантиссы.

Представление вида (1) с мантиссой заданной в виде (2) называется нормализованным представлением чисел в формате с плавающей точкой. Процесс получения нормализованных чисел называется нормализацией.

# Формат с плавающей точкой

Знаковый бит равен 0 для положительных чисел и 1 для отрицательных.

Поскольку поле экспоненты составляет 8 битов, то в нём можно представить значения в диапазоне от -128 до 127.

Нет необходимости хранения бита $b_0$, т.к. он всегда равен единице и является скрытым битом.

При представлении вещественных чисел в формате с плавающей точкой, не имеющих конечного двоичного представления, лишние разряды отбрасываются, например, для чисел с плавающей точкой длиной 32 бит, число

$$1/10 = (0.0001100110\ \ 011\ \ldots)_2$$

# Формат с плавающей точкой

Пусть значение экспоненты принадлежит диапазону $e_{\min} \le e < e_{\max}$ Число называется *представимым* в формате с плавающей точкой, если его можно представить в виде:

$$(-1)^s \cdot m \cdot q^e, \text{ с } e_{\min} \le e < e_{\max}$$

Пусть рассматривается случай 1, когда мантисса удовлетворяет неравенству $q^{-1} \le m < 1$, тогда минимальное представимое число равно $q^{e\min - 1}$ и максимальное $q^{e\max}(1 - q^{-n})$.

# Нарушение законов алгебры

- No associative property for floats
- $(a + b) + (c + d)$ (parallel) $\neq ((a + b) + c) + d$ (serial)
- Looks like a “wrong answer”

## *ПРИМЕР ЗАДАЧИ, ИМЕЮЩЕЙ РЕЗКИЙ РОСТ ОШИБОК ОКРУГЛЕНИЯ*

**Матрица Гильберта** $A = \{a_{ij}\},\ a_{ij} = \dfrac{1}{i+j-1}$

**Обращение матрицы Гильберта порядка 3**

С точностью 2 знака после запятой

$$A = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix}$$

$$A_{прибл}^{-1} = \begin{bmatrix} -1{,}17 & 19{,}51 & -23 \\ 19{,}51 & -112{,}94 & 112 \\ -23 & 112 & -100 \end{bmatrix}$$

Макс. относ. погрешн. более 100%.

С точностью 3 знака после

запятой

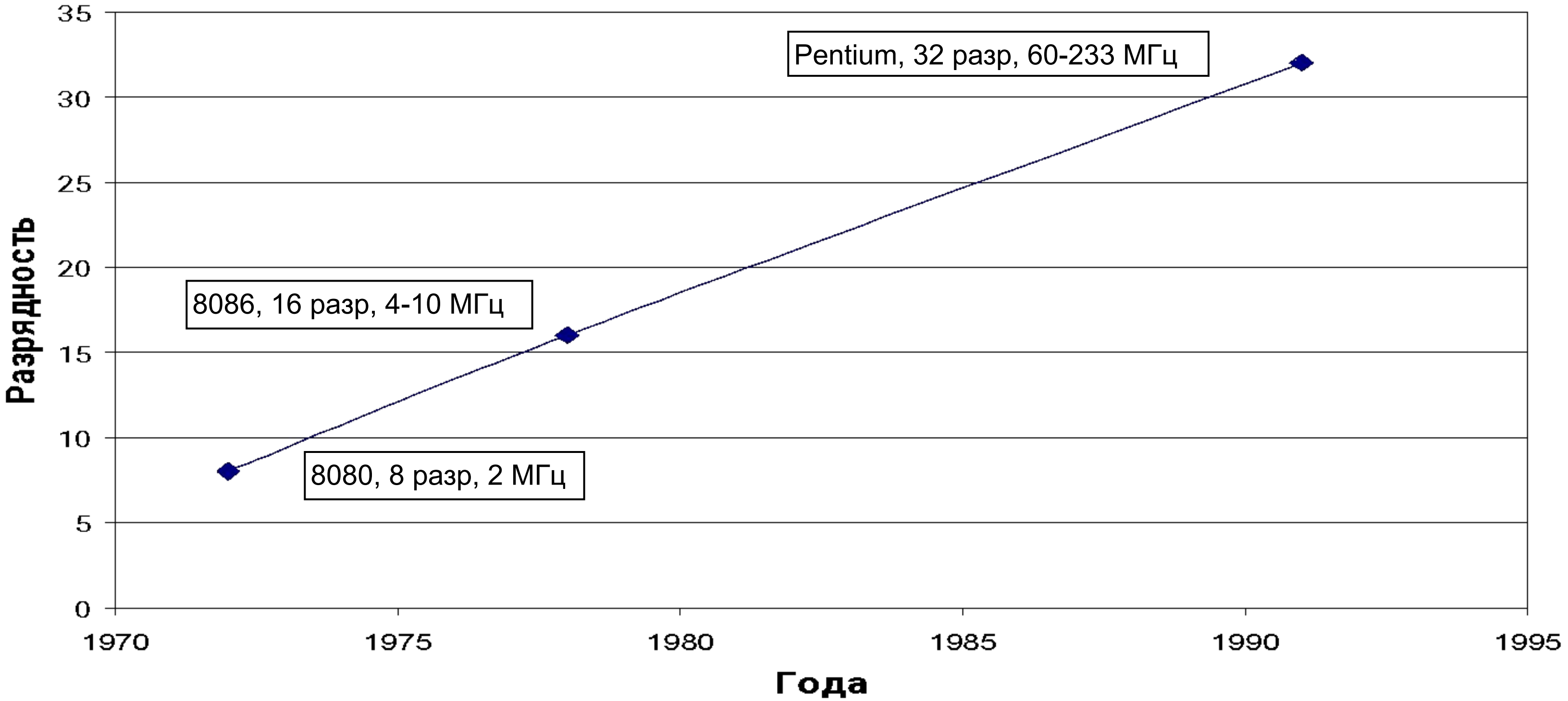
$$A_{прибл}^{-1} = \begin{bmatrix} 10{,}101 & 29{,}598 & 64{,}798 \\ -41{,}039 & -192{,}78 & -202{,}4 \\ 34{,}6 & 202{,}4 & 200 \end{bmatrix}$$

Макс. относ. погрешность более 100%.

**Точный результат:**

$$A_{точн}^{-1} = \begin{bmatrix} 9 & -36 & 30 \\ -36 & 192 & -180 \\ 30 & -180 & 180 \end{bmatrix}$$

# Рост разрядности и тактовой частоты процессоров по годам




**Гипотеза: Технологические трудности создания процессоров высокой разрядности**

## Пример нарушения алгебраического свойства ассоциативности

$(a \oplus b) \oplus c \neq a \oplus (b \oplus c)$ $\oplus -$ сложение чисел с плавающей точкой

Любое число с плавающей точкой в нормальной форме можно представить в следующем виде:

$a \cdot q^b,$

где

$q$ – основание системы счисления,

$b$ – порядок числа,

$a$ – мантисса числа и $q^{-1} \leq |a| < 1$.

$q = 2,$

| Мантисса числа с плав. точкой | Порядок числа | | | |
|---|---|---|---|---|
| | b = 0 | b = 1 | b = 2 | b = 3 |
| 100 | 1/2 | 1 | 2 | 4 |
| 101 | 5/8 | 5/4 | 5/2 | 5 |
| 110 | 3/4 | 3/2 | 3 | 6 |
| 111 | 7/8 | 7/4 | 7/2 | 7 |

## Задачи

1.Доказать, что

$$\left(\frac{1}{2}\oplus\frac{1}{2}\right)\oplus 6 \neq \frac{1}{2}\oplus\left(6\oplus\frac{1}{2}\right)$$

2.Найти диапазон представления чисел с плавающей точкой