Операции над числами с плавающей точкой

Представление вещественных чисел

Представление вещественных чисел

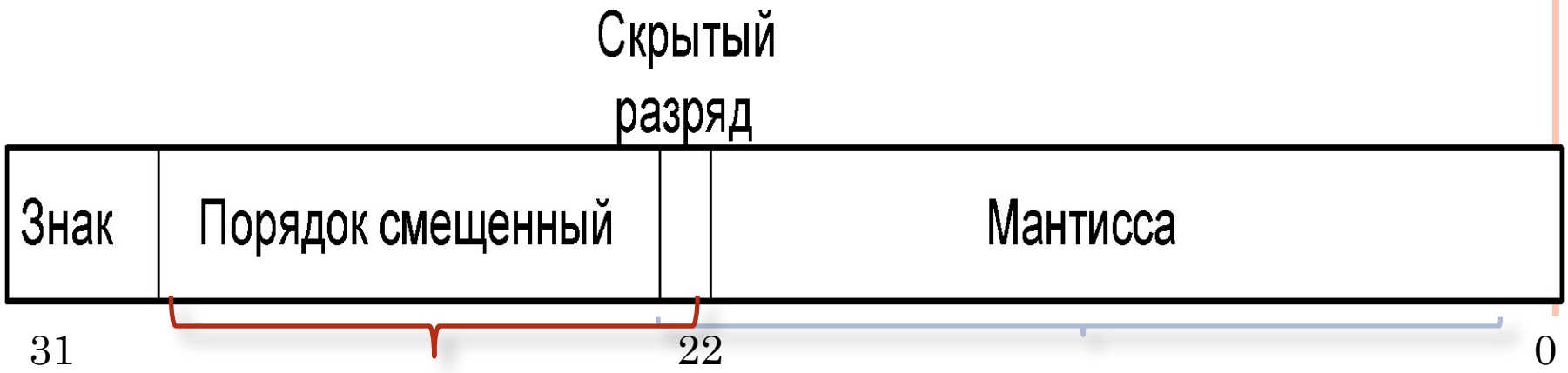
IEEE 754 (IEC 60559) — широко используемый стандарт IEEE, описывающий формат представления чисел с плавающей точкой. Используется в программных (компиляторы с разных языков программирования) и аппаратных (CPU и FPU) реализациях арифметических действий (математических операций).

$$X = \pm S^{Px} q_x,$$

- *определяются требования к порядку- смещенный* $P_{смещ.} = (2^{k-1}-1)+P_{исх.}$

k- кол-во разрядов выделенных под порядок

- *определяются требования к мантиссе-нормализованная* $2 > |q\ x| \geq 1$



Представление вещественных чисел

- Базируется на экспоненциальной форме записи числа:

$$A = m * s^p$$

m - мантисса числа
s – основание СС
p – порядок числа

$$100_{10с.с.} = 0.1*10^3 = 10000*10^{-2}$$

1. определяются требования к порядку: может быть как + , так и -
2. определяются требования к мантиссе.

- Для единообразия представления чисел используется нормализованная форма:

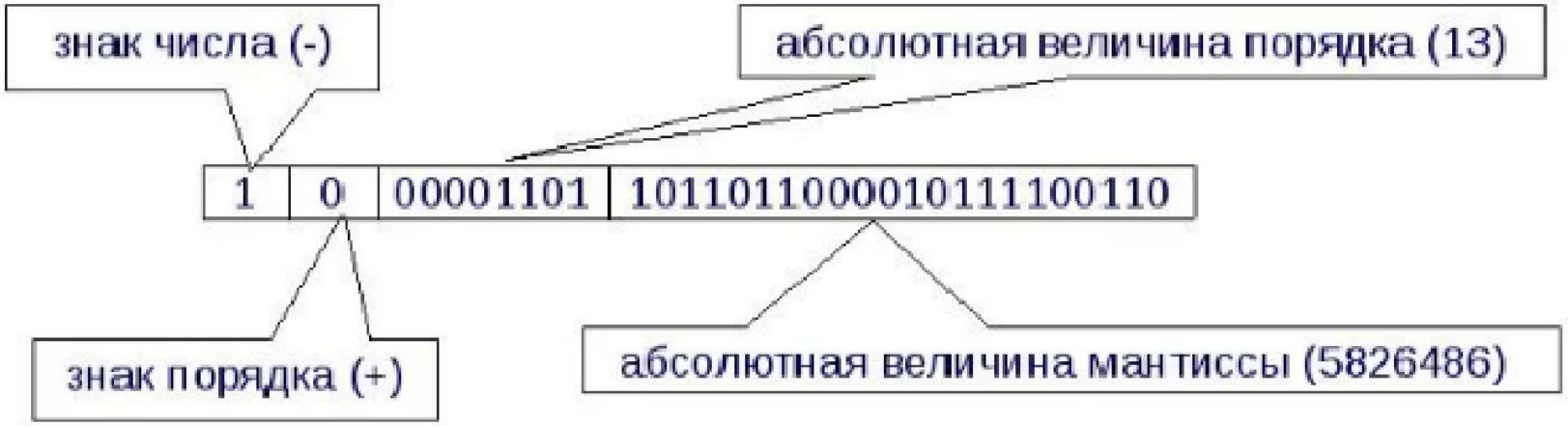
- *$1/s <= |m| < 1$ (правильная дробь и после запятой цифра, отличная от нуля.)*

- **Пример. Преобразуйте число 555,55, записанное в естественной форме, в экспоненциальную форму с нормализованной мантиссой: $555,55 = 0,55555 * 10^3$**

Нормализованная мантисса: 0,55555

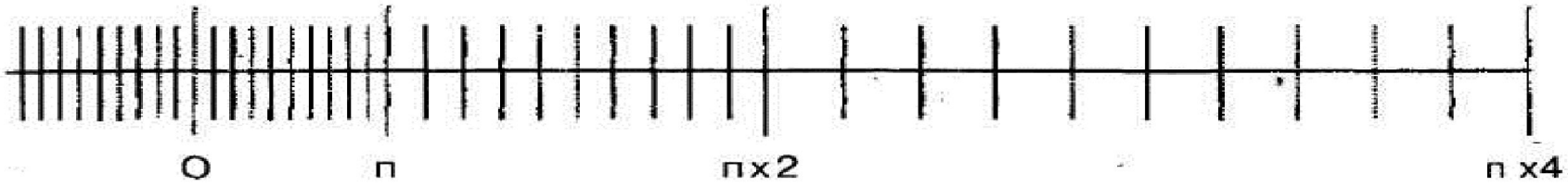
Порядок: p = 3

При представлении чисел с плавающей точкой в разрядах ячейки отводится место для знака числа, знака порядка, абсолютной величины порядка, абсолютной величины мантиссы.



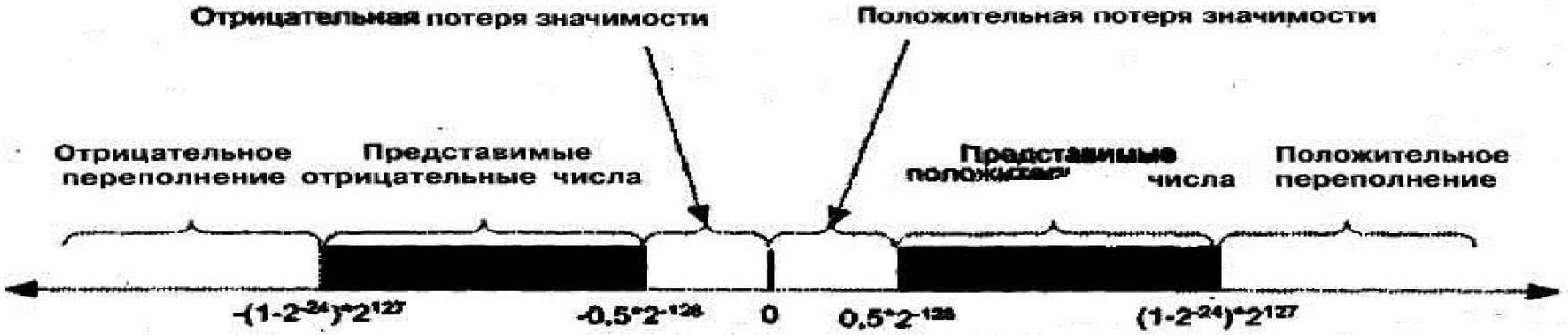
ячейке записано отрицательное двоичное число –1011011000010.11110011

десятичном представлении это будет число –5826.486



Плотность чисел с плавающей запятой на числовой оси

С фиксированной точкой



С плавающей точкой

В процессорах Intel (стандарт IEEE 754) применяется 3 (основных) формата с плавающей точкой: короткий, длинный и расширенный.

	Короткий	Длинный	Расширенный
Разрядность числа	32	64	80
Разрядность мантиссы	24	53	64
Диапазон значений	$10^{-38}…10^{+38}$	$10^{-308}…10^{+308}$	$10^{-4932}…10^{+4932}$
Размерность порядка	8	11	15
Значение $(2^{k-1}-1)$	+127	+1023	+16383
Диапазон $p_{смещ}$	0…255	0…2047	0…32767
Диапазон порядков	-126...+127	-1022+1023	-16382+16383

замечания

- *Рсм*, содержащий во всех разрядах 1 не используется, т.к. зарезервирован для указания на переполнение порядка или потерю значимости мантиссы.
- При этом + и - переполнение идентифицируются соответственно + и -мантиссами, содержащими в обоих случаях 1 во всех цифровых разрядах. Указанием на потерю значимости служит отрицательность порядка мантиссы с 0 во всех цифровых разрядах.
- При нулевых кодах порядка и мантиссы представляемое число полагается равным 0.
- Нулевому порядку в коротком формате соответствует значение фиксированного смещения равное 127, а в разрядной сетке запишется двоичное представление 01111111
- Отрицательному порядку -1, соответствует 127-1=126→ 01111110 Положительному порядку +1, соответствует 127+1=128→10000000,т.е. все положительные порядки имеют старший бит порядка равный 1, а отрицательные-0.

Так как нормализованное число в старшем разряде всегда содержит 1, то при его представлении в памяти появляется возможность считать 1-й разряд вещественного числа единичным по умолчанию. И учитывать его наличие только на аппаратном уровне. Это дает возможность увеличить диапазон представимых чисел.
Это утверждение справедливо только для короткого и длинного форматов.

45.56_{10} **=1011 01**.10 0011 1101 0111 0000 1010 0011 1101 0111 0000 1010 1110 0001 0111 0001

Смещенный порядок 127+5=132 или 10000100

Запись числа в коротком формате со скрытым старшим разрядом мантиссы.
0 10000100 01101100011110101110001

Запись результата в 16-ричной с.с. **42363D71** H

Пример . Записать представление числа 0,089 в форме с плавающей точкой.

Решение

1. Переведем в двоичную СС:

$$0{,}089_{10} = 0{,}\ 000101100000000\ldots0_2$$

2. Запишем в форме нормализованного двоичного числа: $0{,}10110000000000000000 * 10_2^{100}$
3. Вычислим машинный порядок в двоичной СС:

$$P_{СМ} = 01111111\text{-}100 = 01111011$$

3. Запишем число в коротком формате:

0 01111011 01100000000000000000

Шестнадцатеричная форма 3DB00000

Алгоритм записи внутреннего представления вещественного числа

1. Перевести модуль числа в двоичную СС с 24 значащими цифрами.
2. Нормализовать двоичное число.
3. Найти машинный порядок в двоичной СС.
4. Учитывая знак числа, записать его в 4-х байтовом машинном слове.

Пример 1 . Записать представление числа 122,1875 в форме с плавающей точкой в коротком формате

1. Переведем в двоичную СС:
   $122{,}1875_{10} = 1111010,\ 0011000000000000_2$

2. Запишем в форме нормализованного двоичного числа: $0,1111010\ 0011000000000000 * 10^6$

3. Вычислим Рсм :

$$Рсм = 127 + 6 \rightarrow 10000101_{2С.С}$$

3. Запишем число в 4-х байтовой ячейке:

0 10000101 111010 00110000 00000000

Шестнадцатеричная форма 42F46000

плавающей точкой C4011000 восстановить число.

1. Перейдем к двоичному представлению числа:

1100 0100 0000 0001 0001 0000 0000 0000

1 10001000 00000010001 00000000 00000000

восстанавливаем 1 в мантиссе

1 10001000 10000000010001 00000000 00000000

2. В старшем разряде с номером 31 записана 1, значит получен код отрицательного числа. Получим порядок числа:

$$p_{исх}=10001000_2 - 01111111_2=1001_2=9_{10}$$

3. Запишем в виде нормализованного дв. числа с плавающей точкой с учетом знака числа: $-0,10000001000100000000000 \cdot 2^{1001}$
4. В двоичной системе СС число имеет вид: $-100000010,001_2$
5. Переведем число в десятичную СС:

$$-100000010,001_2=-(1*2^8+1*2^1+1*^2-3)=-258,125_{10}$$

Представление чисел в формате с плавающей точкой

- Занимает в памяти ПК 4 *(обычная точность)* или 8 байтов *(двойная точность)*
- Выделяются разряды для хранения знака мантиссы, знака порядка, порядка и мантиссы.
- Максимальное значение порядка числа:

$$1111111^2 = 127_{10}$$

- Максимальное значение числа составляет:

$$2^{127} = 1,\ 70141183460469231731687303715884 * 10^{38}$$

- Максимальное значение положительной мантиссы равно:

$$2^{23} - 1 \sim 2^{23} = 2^{(10*2,3)} \sim 1000^{2,3} = 10^{(2,3*3)} \sim 10^7$$

Максимальное значение чисел обычной точности вычислений составляет

$$1,701411 * 10^{38}$$

Операции сложения над числами с плавающей точкой:

$Z=X+Y= S^{Px} (q_x+q_y/ S^{Px-Py}) = S^{Pz}q_z$

Алгоритм сложения чисел с плавающей точкой:

1.Производится выравнивание порядков чисел. Порядок меньшего (по модулю) числа принимается равным порядку большего, а мантисса меньшего сдвигается вправо на количество разрядов, равное разности порядков.

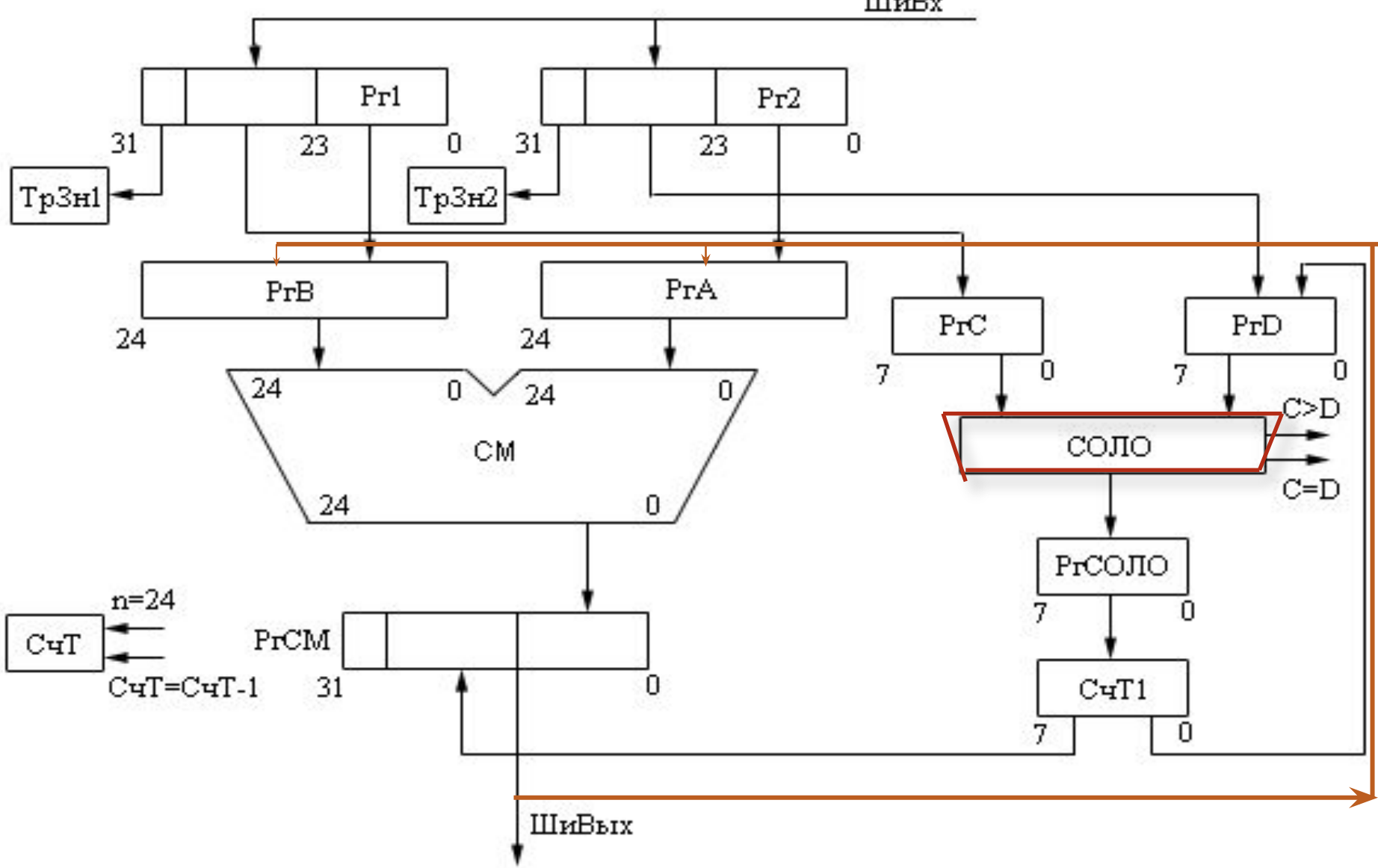
2.Производится сложение (вычитание) мантисс по правилам двоичной арифметики.

3.Нормализация результата.

При сравнении порядков возможны пять случаев:

1. *Px—Py>m* (m — число разрядов мантиссы). В качестве результата суммирования сразу может быть взято первое слагаемое, так как при выравнивании порядков все разряды мантиссы второго слагаемого принимают нулевое значение;
2. *Py—Px>m*- В качестве результата суммирования может быть взято второе слагаемое;
3. *Px—Py = 0*. Можно приступить к суммированию мантисс;
4. *Px—Py= k1(k1<m)*. Мантисса второго слагаемого сдвигается на *k*1 разрядов вправо, затем производится суммирование мантисс;
5. *Py—Px> = k2(k2<m)*. Перед выполнением суммирования мантисс производится сдвиг на *k2* разрядов вправо мантиссы первого слагаемого.

Сложение (вычитание) мантисс производится по правилам сложения (вычитания) чисел с фиксированной точкой.


ШиВх
Рг1
Рг2
31
23
0
31
23
0
Тр3н1
Тр3н2
РгВ
РгА
РгС
РгD
24
24
7
0
7
0
24
0
24
0
СМ
C>D
СОЛО
C=D
24
0
РгСОЛО
7
0
n=24
СчТ
РгСМ
СчТ=СчТ-1
31
0
СчТ1
7
0
ШиВых

Нормализация суммы (разности) производится в случае невыполнения условия $1>qz\geq 1/s$, при этом,

-если $qz\geq X$, Pz увеличивается на 1, а мантисса qz сдвигается на один S-ичный разряд вправо, что дает $|qz|<1$.

-если $|qz|<1$, то мантисса результата сдвигается на разряд влево при одновременном уменьшении порядка результата на 1. Эти операции производятся до тех пор, пока не станет выполняться условие $qz\geq 1/s$.

(При qz = 0 нормализация не выполняется.)

При получении порядка +pz, переполняющего разрядную сетку, должен формироваться сигнал прерывания из-за переполнения порядка.

При получении порядка -pz , переполняющего разрядную сетку, формируются нулевой результат и признак исчезновения порядка.

Пример.

Сложить 4,63 +4,63 =9,26

0.100101000010000…….0 Рсм=127+2=129
+ 0.100101000010000……..0
1.001010000100000…….0

Возникло переполнение:
-необходимо произвести сдвиг вправо на 1 разряд,
-порядок увеличить на 1.

Рсм=130 , q= 0.100101000010000….0

Переведем число в десятичную СС:

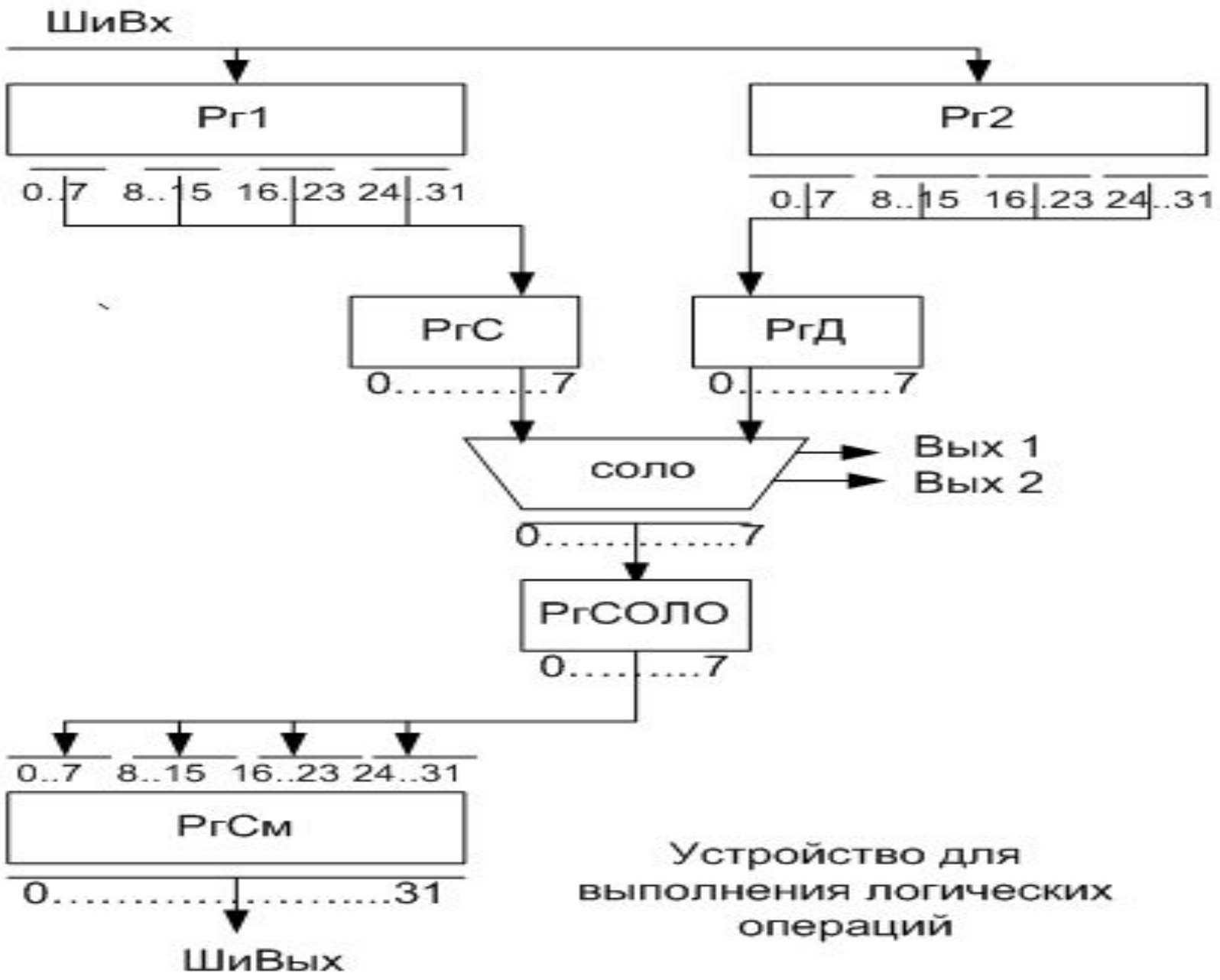
$$1001{,}0100001000 = -(1*2^{0}+1*2^{3}+1*2^{-2}+1*2^{-7}) = 9{,}257812$$

Умножение чисел с плавающей точкой выполняется в соответствии с формулой

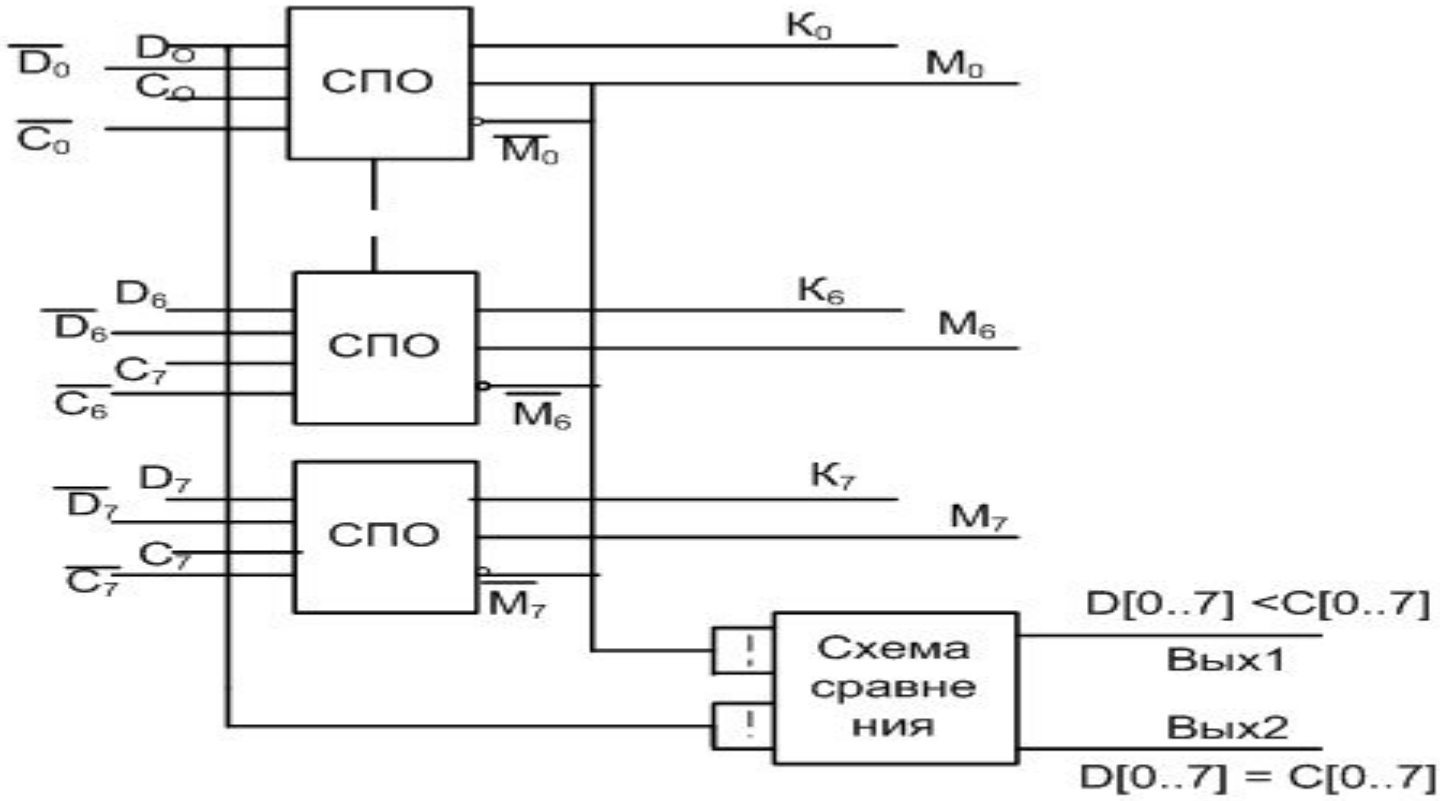
$$Z=X*Y=S^{Px}\,q_x\,*S^{Py}q_y,=S^{(Px\,+\,Py)}q_x\,*q_y=S^{Pz}\,q_z$$

Деление чисел с плавающей точкой выполняется в соответствии с формулой

$$Z=X/Y=S^{Px}q_x/S^{Py}q_y=S^{(Px-Py)}q_x/q_y=S^{Pz}q_z$$



Устройство для выполнения логических операций



Комбинационная схема поразрядной обработки

Вых 1	Вых 2	Результат сравнения
1	1	D<C
0	1	D>C
0	0	D=C

На выходе Ki образуется конъюнкция

$$K_i = \overline{\overline{C}_i \vee \overline{D}_i} = C_i D_i$$

На выходе Mi формируется значение суммы по модулю 2

$$M_i = \overline{\overline{C}_i \overline{D}_i \vee C_i D_i} = C_i \oplus D_i$$

Выходы Ki и Mi соединены со входами РгСОЛО

$\text{РгСОЛО} := \text{РгС} \wedge \text{РгD}$ В РгСОЛО заносится значение состояния выходов Ki

$\text{РгСОЛО} := \text{РгС} \oplus \text{РгD}$ В РгСОЛО заносится значение состояния выходов Mi

Если выполняются 2 микрооперации, то в РгСОЛО заносится результат поразрядной операции ИЛИ

$$\text{РгСОЛО} := \text{РгС} \vee \text{РгD}$$

Задание 1.

Преобразовать следующие числа с плавающей точкой одинарной точности из шестнадцатеричной в десятичную систему счисления.

Таблица 14. Исходные данные

№ вар	Исх.данные	№вар.	Исх.данные
1	4175C28FH	9	41A2041

Задание 2.

Преобразовать следующие числа в формат стандарта IEEE с одинарной точностью. Результаты представить в восьми шестнадцатеричных разрядах.

Таблица 15.Исходные данные

№ варианта	Исх. данные	№ варианта	Исх. данные	№ варианта	Исх. данные
1	9	9	-12.45	17	-9.12
2	5.3210	6.2518	2.86		

Задание 3.

Сложить два числа с плавающей точкой на сумматоре прямого кода. Нормализовать результат и выразить его в шестнадцатеричной системе счисления.

Таблица 16.Исходные данные

№вар.	Исходные данные	№вар.	Исходные данные
1	3EE00000H+18800000H	13	FCFF0000H+ F79A8000H

Задание 4.

Выполнить умножение A*B и деление C/D, сложить полученные результаты, записать их в коротком формате.

Таблица 16. Исходные данные

№ вар	A	B	C	D	№ вар	A	B	C	D
1	15.360	0.38	1.78	132.7	13	201	1.220	9.9	0.49
2	0.57	195.3	0.67	0.89	14	56.37	42.81	0.9	3.57

№ вар.	операция	Условия нормализации мантиссы , особенности структуры, код
1	Сложение	$1>=\|q_x\|>2$, прямой код, схема СОЛО
2	Вычитание	1>=|qx| >2, прямой код, схема СОЛО
3	Умножение	$1>=\|q_x\|>2$, прямой код со старших разрядов
4	Деление	$1>=\|q_x\|>2$, доп. код без восстановления остатка.
5	Сложение	1>=|qx| >2, прямой код , использовать Σ для выравнивания порядков
6	Вычитание	$1/s>\|q_x\| \geq 1$, прямой код , использовать Σ для выравнивания порядков
7	Деление	$1>=\|q_x\|>2$, прямой код, ускоренный алгоритм
8	Сложение	1>=|qx| >2, доп. код, схема СОЛО
9	Деление	$1>=\|q_x\|>2$, прямой код с восстановл. остатка
		$1>=\|q_x\|>2$, доп. код, схема СОЛО
10	Сложение	$1>=\|q_x\|>2$, прямой код, схема СОЛО
11	Вычитание	$1/s>\|q_x\| \geq 1$, доп. код , использовать Σ для выравнивания порядков
12	Умножение	$1/s>\|q_x\| \geq 1$., доп.код
13	Вычитание	$1/s>\|q_x\| \geq 1$, прямой код , использовать Σ для выравнивания порядков
14	Умножение	1>=|qx| >2, прямой код, М.С со старших разрядов
15	Деление	$1/s>\|q_x\| \geq 1$., доп.код
16	Умножение	1>=|qx| >2, прямой код, М.С со старших разрядов
17	Сложение	$1>=\|q_x\|>2$, прямой код, использовать Σ для выравнивания порядков
18	Сложение	$1/s>\|q_x\| \geq 1$, доп.код.,использовать схему СОЛО