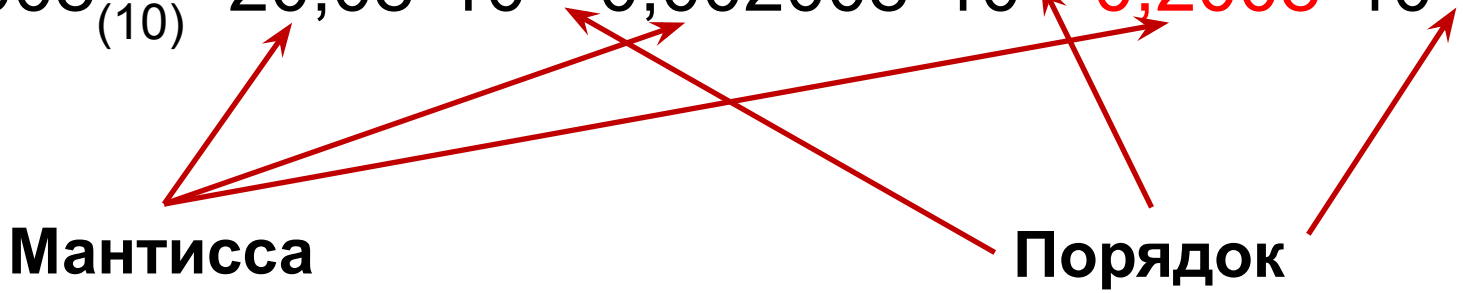
Формат с плавающей точкой (float)

Одной из форм записи вещественных чисел является их представление в ***экспоненциальном виде***, в котором отдельно записывают мантиссу числа и порядок числа.

Пример экспоненциальной формы числа $2008_{(10)}$:

$2008_{(10)}=20,08*10^{2}=0,002008*10^{6}=0,2008*10^{4}$



Мантисса равна 20,08 или 0,002008, или 0,2008.
Порядок равен соответственно 2 или 6, или 4.

Любое число в экспоненциальной форме имеет множество представлений.

$1=0{,}00001*10^{5}=1000*10^{-3}$ и т.д.

Среди этих представлений выделили *нормализованное представление числа*:

$2008_{(10)}=0{,}2008*10^{4}$

Для каждого числа это представление – *единственное*.

$1=0{,}1*10^{1}$

При нормализированном представлении числа в экспоненциальной форме мантисса (М) должна быть в интервале

$0{,}1_{(d)} \le M < 1_{(d)}$, где d – основание системы счисления.

Операнды в цифровом процессоре в формате с плавающей точкой (ПТ) - float представляют числа в *экспоненциальной форме*.

Такой формат чисел в компьютерах используется в для научно-технических расчетов, когда в вычислениях диапазон чисел может варьироваться от очень *малых величин* до *очень больших*, т.е. нужно обеспечить большой диапазон вычислений. Это плюс.

Но в отличие от формата с ФТ, в котором выполняются абсолютно точные вычисления, операции в таком формате выполняются с приближением, определяемым разрядной сеткой процессора.

В истории IT- технологий существовало много форматов чисел в формате с ПТ.

В настоящее время общепринятым стандартом представления операндов в формате с плавающей точкой в цифровом процессоре является стандарт **IEEE 754**.

В IEEE 754:

- мантисса представляется в прямом коде;
- порядок «смещен» (увеличен) на константу.

Смещение порядка на константу позволяет обойтись без явного бита знака порядка. Если значение смещенного порядка больше константы смещения, он – положительный, если меньше – отрицательный. Такое решение позволило реализовать более простые алгоритмы операций над порядками.

Если константа смещения равна 127, то:

$0,1*10^{-3}$ будет записано, как $0,1*10^{-124}$;

$0,2008*10^{4}$ будет записано, как $0,2008*10^{131}$

Есть два формата представления чисел с плавающей точкой стандарта **IEEE 754** в оперативной памяти процессора :

- «короткое вещественное (КВ)»,
- «длинное вещественное (ДВ)».

Они отличаются диапазоном представимых в них чисел.

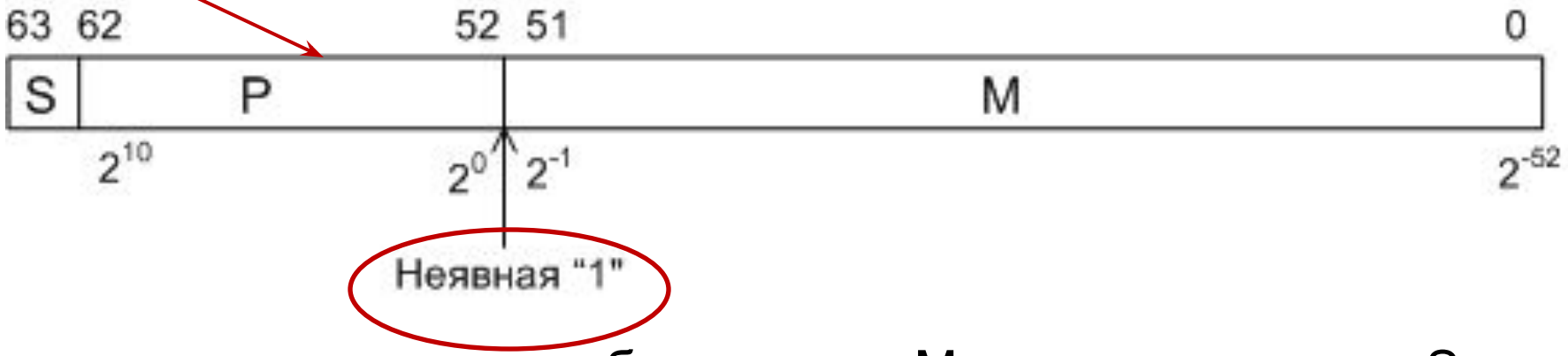
В самом процессоре арифметические операции выполняются всегда в формате:

«внутреннее (расширенное) вещественное (ВВ)».

Т.е. при передаче из оперативной памяти в процессор форматы КВ и ДВ преобразуются в ВВ и наоборот.

КВ (4 байта)

ДВ (8 байт)



Используются следующие обозначения: M – мантисса числа; S – знак мантиссы; P – порядок числа.

В формате КВ под мантиссу отводится 24 бит, а под порядок – 8 бит. Величина порядка операнда смещена на $127_{(10)}$, т.е. $P=P_x+127_{(10)}$.

В формате ДВ под мантиссу отводится 53 бит, а под порядок – 11 бит. Величина порядка операнда смещена на $1023_{(10)}$, т.е. $P=P_x+1023_{(10)}$.

«Скрытый» бит мантиссы

Поскольку при нормализованном представлении операнда ($0,1_{(2)} \le M < 1_{(2)}$) в двоичной системе счисления первая цифра мантиссы после запятой всегда будет равна «1», т.е. 0,1……

это можно использовать *для увеличения диапазона* представимых чисел в *оперативной памяти*, для чего диапазон представления мантиссы нормализованного числа в стандарте IEEE 754 меняется на диапазон $1_{(2)} \le M < 2_{(2)}$.

Причем единица целой части мантиссы учитывается неявно (неявная единица), т.е. под нее не отводится бит. *В таком виде операнд хранится в памяти процессора*. При выполнении арифметических операций над операндом, при его извлечении из памяти в регистр процессора (формат ВВ) этот скрытый бит восстанавливается, т.е. присутствует в явном виде

Алгоритм преобразования вещественного десятичного числа в двоичное число с плавающей точкой формата IEEE 754 (на примере числа = $8{,}125_{(10)}$)

1.Перевести целую часть вещественного числа в двоичную систему и поставить после нее десятичную точку (для заданного примера: 1000).

2. Перевести дробную часть вещественного числа в двоичную систему с точностью для определения значений всех битов мантиссы, предусмотренных форматом (КВ, ДВ, ВВ) (для заданного примера: 0,0010…0)

3. Записать полученное значение дробной части после десятичной точки. Если значение мантиссы меньше выделенного под нее количества разрядов, то дополнить дробную часть незначащими нулями справа до предусмотренного форматом размера. Для заданного примера:
1000,0010000000000000000000000000 (всего 31 бит)

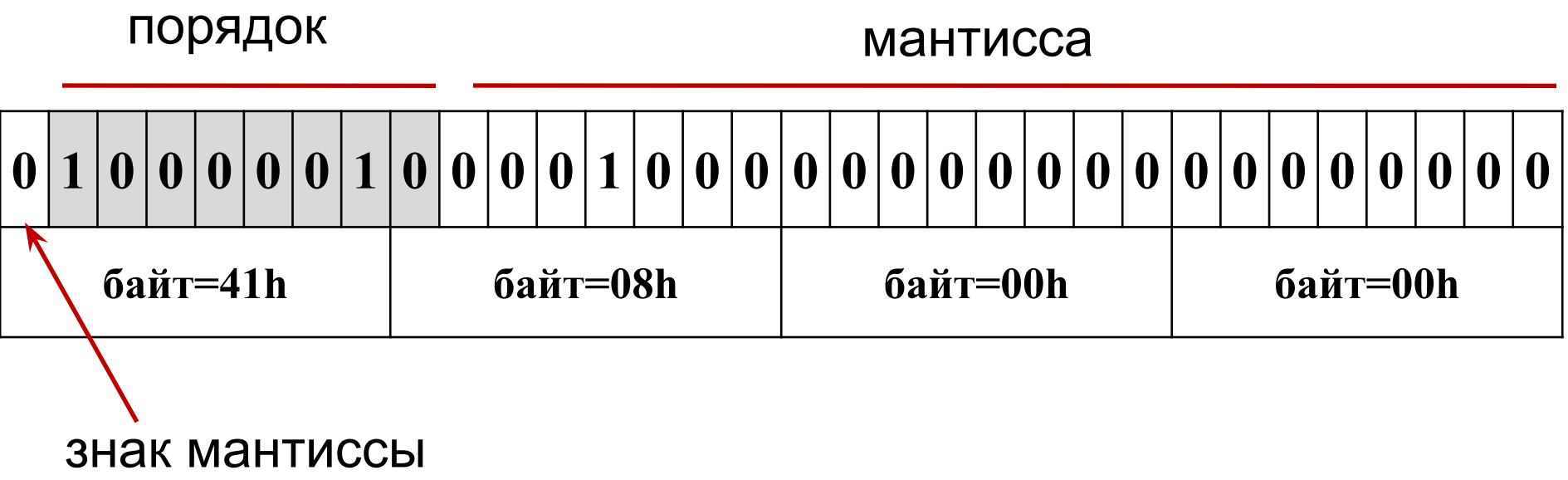
4. Представить число в экспоненциальной форме. Для заданного примера:
$0,\ 1000001000000000000000000000000 * 10^{100}$

5. Нормализовать полученное двоичное число ($1_{(2)} \leq M < 2_{(2)}$), определив тем самым значение порядка. Для заданного примера:
$1{,}00000100000000000000000000000000 * 10^{11}$

6. К порядку прибавить смещение в соответствии с форматом, и представить его в двоичном виде. В результате будет получен смещенный порядок для выбранного формата (для заданного примера: $3+127=130=10000010_{(2)}$).

7. Записать значение порядка и значение мантиссы в соответствующие биты формата КВ или ДВ (у мантиссы, отбрасывается единица целой части).

8. Если число положительное, то в самый старший разряд (знак мантиссы) следует записать 0, если отрицательное, то 1.



Рассмотрим другие примеры представления операндов в формате КВ.

Напоминание:

- мантисса – M ($1_{(2)} \le M < 2_{(2)}$);
- порядок – P (смещен на $127_{(10)}$).

<u>Пример.</u> Представить число $16,AC_{(16)}$ в формате КВ.

Перевод в двоичную систему:

$16,AC_{(16)}=10110,10101100_{(2)}=1,011010101100_{(2)}*10^{100}$.

$M=1,011010101100_{(2)}$;

$P=100_{(2)}+(127_{(10)}=1111111_{(2)})=10000011_{(2)}$.

Тогда формат КВ этого числа (красным цветом выделены биты порядка) будет (целая часть мантиссы «скрыта»):

0	1	0	0	0	0	0	1	1	0	1	1	0	1	0	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0
байт=41h									**байт=0B5h**								**байт=60h**								**байт=00h**							

Еще пример:

Представить число $-7,C8_{(16)}$ в формате КВ.

Перевод:

$-7,C8_{(16)} = -111,11001000_{(2)} = -1,1111001000_{(2)} * 10^{10}$.

$M = -1,1111001000_{(2)}$;

$P = 10_{(2)} + (127_{(10)} = 1111111_{(2)}) = 10000001_{(2)}$.

Тогда формат КВ этого числа (красным цветом выделены биты порядка) будет(целая часть мантиссы «скрыта»):

1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
байт									**байт**								**байт**								**байт**							

Еще пример:

Представить число $-0{,}0C8_{(16)}$ в формате КВ.

Перевод:

$$-0{,}0C8_{(16)} = -0{,}000011001000_{(2)} = -1{,}1001000_{(2)} * 10^{-101}.$$

$M = -1{,}1001000_{(2)}$;

$P = -101_{(2)} + (127_{(10)} = 1111111_{(2)}) = 01111010_{(2)}$.

Тогда формат КВ этого числа (красным цветом выделены биты порядка) будет(целая часть мантиссы «скрыта»):

1	0	1	1	1	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
байт									байт								байт								байт							