Представление чисел с плавающей точкой

Числа в цифровых устройствах представляются с фиксированной или плавающей точкой. Настоящая лекция посвящена представлению чисел с плавающей точкой, фиксированная точка рассматривается в предыдущей лекции.

Экспоненциальная[1] (стандартная) форма записи числа

Т.к. представление чисел с плавающей точкой основывается на экспоненциальной (или стандартной) записи, рассмотрим сначала ее.

При экспоненциальной записи число представляется в виде

$$A = M \cdot B^p, \quad (1)$$

где M – мантисса[2], B – основание степени, p – порядок числа. Иногда порядок называют показателем степени или экспонентой. Например, запись постоянной Больцмана в экспоненциальном виде выглядит так: $1{,}38064852 \times 10^{-23}$. Здесь 1,38064852 – мантисса, 10 – основание степени, –23 – порядок. Часто для удобства оформления или отображения вместо записи B^p указывают только символ E: 1,38064852E–23. При этом полагается, что значение основания понятно из контекста (чаще всего 10).

Одно и то же число может быть записано в виде (1) разными способами. Например, число 123,456 в экспоненциальном виде может быть представлено, например, в виде $1{,}23456 \times 10^2$, $12{,}3456 \times 10^1$, $123{,}456 \times 10^0$, $1234{,}56 \times 10^{-1}$ и др. Для устранения этих неоднозначностей отдельно выделяют **нормальную**[3] и **нормализованную** формы числа.

В нормальной форме порядок p выбирается таким образом, чтобы $0 \le \mathrm{abs}(M) < 1$. При этом сохраняется описанная выше возможность неоднозначного представления чисел, которые начинаются с нуля. Например, $0{,}0001 \times 10^3$, $0{,}001 \times 10^2$, $0{,}01 \times 10^1$ и др.

В нормализованной форме порядок p выбирается таким образом, чтобы $1 \le \mathrm{abs}(M) < B$. При этом если основание степени равно 2, то старший разряд мантиссы будет всегда равен 1. При нормализованной записи разделитель целой и дробной частей мантиссы (точка или запятая) всегда будет находиться между двумя старшими значащими (т.е. не равными нулю) цифрами в позиционной записи числа. Т.е. точка всегда помещается после первой слева значащей цифры.

Нормализованная запись удобна для выполнения операций сравнения и некоторых математических операций. Также она устраняет неоднозначность представления в виде (1).

Недостатком нормализованной записи является невозможность записи нуля, поэтому для его отображения используется специальный признак (см. ниже).

Запись числа с плавающей точкой

Запись числа с плавающей точкой есть не что иное, как машинная реализация стандартной формы представления чисел.

При представлении с плавающей точкой число характеризуется знаком, мантиссой и порядком. Сначала записывается знак числа, потом – порядок и далее – мантисса. Для записи числа отводится $N = Np + Nm + 1$ битов, самый левый знаковый разряд – старший (most significant bit, MSB), самый правый – младший (least significant bit, LSB) (рис. 1).

---

[1] В англоязычной литературе применяется термин scientific notation, что дословно переводится, как “научный” формат.
[2] В англоязычной литературе используется термин significand, термин mantissa является устаревшим.
[3] Иногда нормальную форму называют инженерной нормализованной.

Знак	Порядок				Мантисса			
	$Np - 1$	…	1	0	$Nm - 1$	…	1	0
S	p_{Np-1}	…	p_1	p_0	d_{Nm-1}	…	d_1	d_0
MSB								**LSB**

Рис. 1. Представление числа с плавающей точкой

В математическом виде число X с плавающей точкой записывается в следующем виде:

$$X = (-1)^S \cdot M_n \times B^p, \tag{2.a}$$

или

$$X = (-1)^S \cdot \frac{M}{B^{N_M - 1}} \cdot B^p \tag{2.б}$$

где S – знак числа, $S = 0$ для положительных чисел и $S = 1$ для отрицательных; M – мантисса, целое неотрицательное число; M_n – мантисса в нормализованной форме (старший разряд 1, после него следует точка); p – порядок, целое число; B – основание системы счисления, целое число, не меньшее 2 (обычно 2 в машинном представлении или 10 в письме).

Числа с плавающей точкой записываются в нормализованной форме. Мантисса содержит все цифры в записи числа и определяет точность его представления, а порядок определяет положение точки. Если порядок положительный, то запятая смещается вправо, если отрицательный – влево. Исходное положение точки, как следует из (2) или определения нормализованной формы, – сразу после первого слева значащего разряда.

Из-за возможных изменений порядка в процессе вычислений положение точки тоже может меняться. Отсюда и возникло название представления: плавающая точка. Обратим внимание на то, что знак × в записи (2.a) – это часть записи, показывающий основание системы счисления и порядок (т.е. на сколько необходимо сдвинуть точку), а не символ умножения[4].

Например, число 1234,56789 при записи с плавающей точкой в десятичной системе счисления представляется мантиссой 123456789 и порядком 3. Для получения исходного числа необходимо в мантиссе поместить точку/запятую после первой цифры 1 и сдвинуть ее на 3 разряда вправо.

1. Записать мантиссу: 123456789.
2. Разместить запятую после первой слева цифры: 1,23456789.
3. Сдвинуть запятую на три разряда вправо: 1*234*,56789

Число 0,01234,56789 будет представлено такой же мантиссой, как и предыдущее число, но порядок будет равен –2.

Рассмотрим теперь пример записи числа в двоичной системе счисления. Пусть исходное (неотрицательное) число 1101 1110, 0010 1101. Тогда с плавающей точкой оно будет представлено мантиссой 1101 1110 0010 1101 и порядком 111 (7_{dec}).

При представлении чисел с плавающей точкой в нормализованной форме в двоичной системе счисления старший разряд мантиссы всегда равен 1. Это значит, что можно экономить один разряд в записи числа, помня при выполнении математических операций о том, что в нем всегда 1.

Стандарт IEEE 754

Изначально не существовало единого стандарта для записи чисел с фиксированной точкой, т.е. правила выполнения арифметических операций, округления, представления нуля и бесконечности варьировались от одной программы или архитектуры компьютера к другой. Для

[4] Это связано с тем, что в общем случае, как видно из того же стандарта IEEE 754, при выполнении операций умножения возможны ситуации, которые можно отрабатывать по-разному (например, обработка чисел типа NaN (not a number), ±∞). Т.е. сама операция умножения в общем случае требует стандартизации.

преодоления этих проблем в 70-х годах 20-го века начались работы по стандартизации представления чисел в плавающей точке и операций над ними. В результате острой борьбы лобби различных компаний (DEC, Intel, Motorola и др.) родился стандарт IEEE 754 (от Intel).

В стандарте IEEE 754 определены форматы представления чисел с половинной, одинарной, двойной, четырехкратной и восьмикратной точностью. Далее будем рассматривать только представления с основанием степени 2 (табл. 1).

В (2) значение порядка может быть как отрицательным, так и неотрицательным. В стандарте IEEE 754 для записи порядка используются беззнаковые целые числа. Для корректных вычислений для каждой точности определена величина смещения порядка, которую необходимо вычитать из записанного. Например, в записи числа одинарной точности порядок равен $00001111_{bin} = 15_{dec}$. Это значит, что истинное значение порядка, которое должно использоваться в математических операциях равно 15 – 127 = –112.

Важной особенностью при записи мантиссы является тот факт, что в старшем ее разряде в нормализованной форме всегда единица. Это значит, что ее можно не записывать и, тем самым, сэкономить один разряд. В стандарте IEEE 754 старшая единица мантиссы не записывается, но учитывается при математических операциях. Поэтому в табл. 1 число разрядов представлено в виде $x + 1$: x – это число разрядов, физически выделенное для записи мантиссы, а $x + 1$ – полное число разрядов мантиссы.

Напомним, что старший разряд в числе – знаковый.

Таблица 1. Форматы чисел с плавающей точкой в стандарте IEEE 754

Точность	Число бит в записи	Число разрядов в мантиссе	Число разрядов в порядке	Смещение порядка	Минимальный порядок (p_{min})	Максимальный порядок (p_{max})
Половинная	16	10 + 1 = 11	5	15	–14	15
Одинарная	32	23 + 1 = 24	8	127	–126	127
Двойная	64	52 + 1 = 53	11	1023	–1022	1023
Четырехкратная	128	112 + 1 = 113	15	16383	–16382	16383
Восьмикратная	256	236 + 1 = 237	19	262143	-262142	262143

Например, пусть записано следующее число с плавающей точкой с одинарной точностью: 0 01110000 10100000000000000000000.

1. В старшем разряде 0 – число неотрицательное.
2. В следующих 8 разрядах число $01110000_{bin} = 112_{dec}$. Значит истинное значение порядка 112 – 127 = –15.
3. В мантиссе (с учетом неявной старшей единицы) записано число целое число 13 631 488.
4. В соответствии с (2) в результате имеем число $X = (-1)^0 \cdot 13631488 \cdot 2^{-15}/2^{24-1} = 0{,}000495910644531\,25$.

Специальные числа в стандарте IEEE 754

Помимо определения представлений чисел с плавающей точкой с различной точностью в стандарте IEEE 754 определены специальные способы для обозначения нулей, бесконечностей и неопределенностей. Также в стандарте указан порядок отображения денормализованных чисел.

В случае использования нормализованной формы для записи чисел невозможно явно отобразить ноль. Для отображения нуля в IEEE 754 используется представление со всем нулями в мантиссе и всеми нулями в порядке (соответственно, вычисляемый с учетом смещения порядок равен –127). Ноль может быть положительным или отрицательным. Например, положительный и отрицательный нули с одинарной точностью будут равны 0(1) 00000000 00000000000000000000000. Операции с нулем выполняются не так, как обычные операции, что усложняет обработку, но позволяет избежать многих странных моментов при выполнении математических операций.

Для отображения бесконечностей $\pm\infty$ используется мантисса со всеми нулями и порядок со всеми единицами. Так, $\pm\infty$ в числе с половинной точностью будут равны 0(1) 11111 0000000000.

Для отображения неопределенностей, которые могут возникать при выполнении математических операций (например, 0/0), в стандарте используется следующий формат: в порядке все единицы, а в мантиссе – ненулевое число.

Денормализованные числа используются для представления сверхмалых для данного представления чисел. Имеются в виду, числа, модуль которых меньше модуля минимального представимого нормализованного числа. Для записи таких чисел используется порядок со всеми нулями и ненулевая мантисса. При этом полагается, что порядок равен p_{min}, а неявный старший бит мантиссы равен нулю. Так в половинной точности минимальное положительное нормализованное число равно $2^{-14} \approx 6{,}1 \times 10^{-5}$.

Таким образом, числа с плавающей точкой можно записать в следующем виде.

– Нормализованная форма: $(-1)^S \cdot 1.M \times 2^p$, если $p_{min} \leq p < p_{max}$;

– Денормализованная форма: $(-1)^S \cdot 0.M \times 2^{p_{\min}}$, если $p \leq p_{min} - 1$.

Сравнение представлений чисел в фиксированной и плавающей точках

Важным преимуществом записи чисел в формате с плавающей точкой над записью чисел с фиксированной точкой заключается в том, что возможно использование значительно большего диапазона чисел при сохранении относительной[5] погрешности постоянной как для малых, так и для больших чисел. Это связано с тем, что в отличие от чисел с фиксированной точкой сетка отображаемых чисел неравномерная: она гуще вблизи чисел с малыми порядками и редкая для чисел с большими порядками.

Так для плавающей точки половинной точности целые числа между 0 и 2047 представляются точно, между 2048 и 4095 округляются вниз до ближайшего четного числа, от 4096 до 8191 округляются вниз до ближайшего числа, кратного 4, и т.д.

Недостатками представления чисел с плавающей точкой по сравнению с представлением с фиксированной точкой являются более сложная реализация математических операций и, как следствие, меньшее быстродействие и большее энергопотребление. Также необходимо аккуратно выполнять математические операции, если один операнд очень большой, а другой – очень мал. Это связано с неассоциативностью операций сложения и умножения над числами с плавающей точкой (например, от перемены мест слагаемых результат в общем случае меняется). Например, $(10^{30} + 1) - 10^{30} = 0$, а $(10^{30} - 10^{30}) + 1 = 1$.

Таким образом, в случае, когда одновременно важен диапазон обрабатываемых чисел и точность представления необходимо использовать представление с плавающей точкой. А, если необходимы высокое быстродействие или низкое энергопотребление в ущерб точности и диапазону, необходимо воспользоваться записью чисел с фиксированной точкой.

[5] Относительная погрешность измерения – это отношение абсолютной погрешности измерения к значению измеряемой величины.