Zapis pozycyjny

wartość liczby w systemie pozycyjnym ★

$$N_j = qN_{j-1} + a_j$$

Podstawa q jest całkowita i q ≥ 2. Cyfry a_0, a_1, …, a_m czytamy od najbardziej znaczącej do najmniej znaczącej. Każda cyfra a_j jest liczbą całkowitą od 0 do q − 1. Przyjmujemy $N_0 = a_0$, a następnie dla j = 1, …, m obliczamy $N_j = qN_{j-1} + a_j$; wartością całej liczby jest N_m. N_{j-1} oznacza nieujemną całkowitą wartość dotychczas odczytanego prefiksu. W systemie binarnym q = 2, więc każda cyfra ma wartość 0 albo 1.

Liczby binarne

największa liczba bez znaku na n bitach ★

$$N_{\max} = 2^n - 1$$

Liczba bitów n jest całkowita i dodatnia. Każdy bit ma wartość 0 albo 1; zakres liczb bez znaku to od 0 do $2^n - 1$.

liczba różnych kodów na n bitach $K = 2^n$

bity w bajcie $b = 8B$

Liczby ze znakiem i zmiennoprzecinkowe

zakres liczb w kodzie U2 $-2^{n-1} \le x \le 2^{n-1} - 1$

n bitów, n ≥ 1; x całkowite.

wartość słowa w kodzie U2

$$x = -b_{n-1}2^{n-1} + \sum_{i=0}^{n-2} b_i 2^i$$

Bity $b_i \in \{0,1\}$; najbardziej znaczący bit ma ujemną wagę.

normalna liczba binarna IEEE 754

$$x = (-1)^s \left(1 + \frac{F}{2^p}\right) 2^{E-B}$$

Tylko liczby normalne: 0<E<2^e−1; s bit znaku, F całkowite pole ułamka, p jego szerokość. Zera, liczby subnormalne, nieskończoności i NaN mają osobne reguły. e to liczba bitów pola wykładnika E; dla binary32 e = 8, p = 23 i B = 127.

U2: znaczenie bitów i wartość słowa

bit	b_i	waga	wkład do x
b7	1	-128	-128
b6	0	64	0
b5	1	32	32
b4	1	16	16
b3	0	8	0
b2	1	4	4
b1	1	2	2
b0	0	1	0

n = 8; x = −74; zakres: −128 ≤ x ≤ 127.

U2: znaczenie bitów i wartość słowa. Przykład domyślny: słowo 10110110 ma n = 8 bitów. Bit b_7 ma wagę −128, pozostałe wagi są dodatnie. x = −128 + 32 + 16 + 4 + 2 = −74. Zakres 8-bitowego U2: −128 ≤ x ≤ 127.

IEEE 754 binary32: znak, wykładnik i ułamek

pole	bity (od najstarszego)	wartość dziesiętna
s (1 bit)	0	0
E (8 bitów)	10000001	129
F (23 bity)	01000000000000000000000	2097152

x = 5; s = 0, E = 129, F = 2097152; liczba normalna.

IEEE 754 binary32: znak, wykładnik i ułamek. Dla binary32: p = 23, B = 127. Przykład x = 5: s = 0, E = 129, F = 2097152; x = (1 + 1/4)·2². Wzór z początkową jedynką dotyczy wyłącznie 1 ≤ E ≤ 254. E = 0 oznacza zero lub liczbę subnormalną; E = 255 — nieskończoność lub NaN. Podziały linii w polu F nie zmieniają kolejności bitów.

Arytmetyka binarna bez obcinania wyniku

liczba	zapis binarny	dziesiętnie
a	1111	15
b	1	1
wynik	10000	16

1111 + 1 = 10000 (dziesiętnie 16). Bez obcięcia wyniku.

Arytmetyka binarna bez obcinania wyniku. Dane wejściowe są binarnymi liczbami bez znaku. Dodawanie, odejmowanie, mnożenie i dzielenie całkowite działają na liczbach dowolnej potrzebnej długości wyniku. Znak minus przy wyniku nie jest zapisem U2. Przykład domyślny: 1111 + 1 = 10000.

Symbole i jednostki

symbol	znaczenie	jednostka
a_j	kolejna cyfra liczby; cyfry czytamy od najbardziej znaczącej	—
b	liczba bitów	bit
b	liczba bitów jako wartość liczbowa	—
b_i	i-ty bit	
b_{n-1}, s	bit znaku	
B	liczba bajtów	B
B	liczba bajtów jako wartość liczbowa	—
B	obciążenie wykładnika	
E	pole wykładnika	
F	pole ułamka	
i	indeks bitu	
j	indeks kolejnej cyfry (od 1 do m)	—
K	liczba różnych ciągów bitów o długości n	—
n	liczba bitów (liczba całkowita)	—
n	liczba bitów	
$N_{\max}$	największa liczba bez znaku zapisywana na n bitach	—
N_j	wartość prefiksu liczby do cyfry o indeksie j	—
N_{j-1}	wartość prefiksu przed dopisaniem cyfry a_j	—
p	liczba bitów ułamka	
q	podstawa systemu liczbowego	—
x	liczba ze znakiem	