Представление текстовой информации в ЭВМ

Информация:

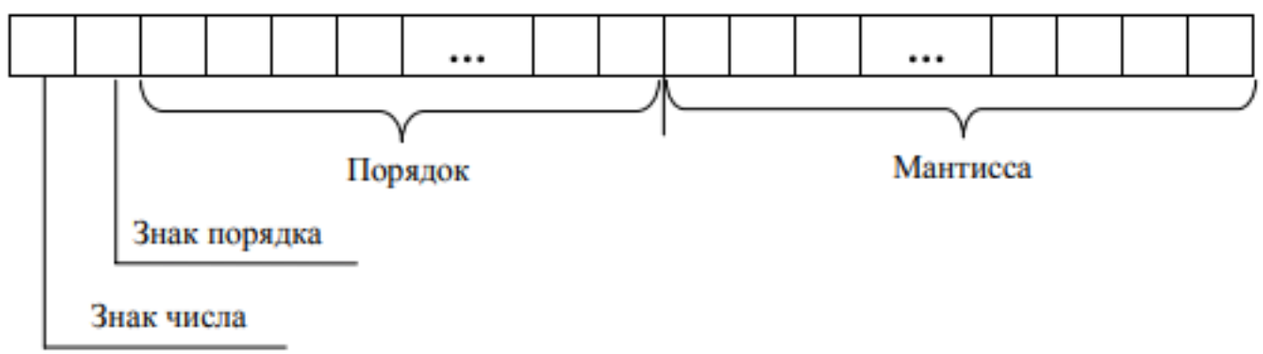
Текстовая
Аудио
Графическая
Числовая
Видео

Двоичный код:

00110101

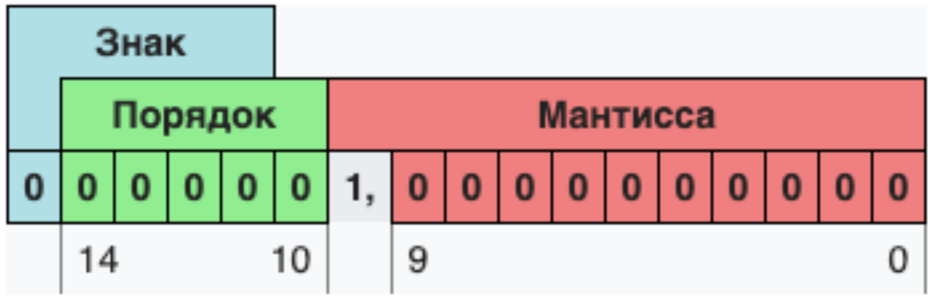
Единица информации в компьютере - один бит, т.е. двоичный разряд, который может принимать значение 0 или 1
В одном байте можно закодировать значение одного символа из 256 возможных (2^8).

1Кбайт = 2^10
1Мбайт = 1024 Кбайт
И тд

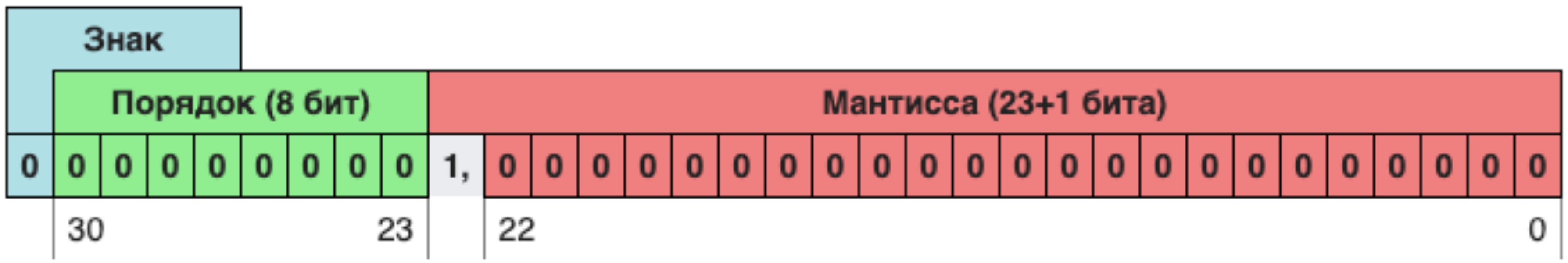


Представление числа с плавающей точкой в виде набора битов

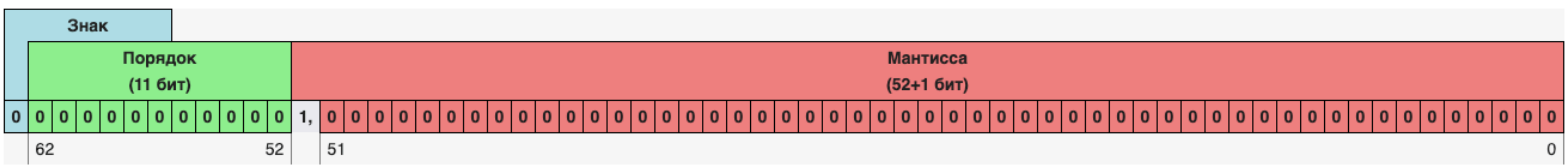
Число половинной точности (Binary16, Half precision)



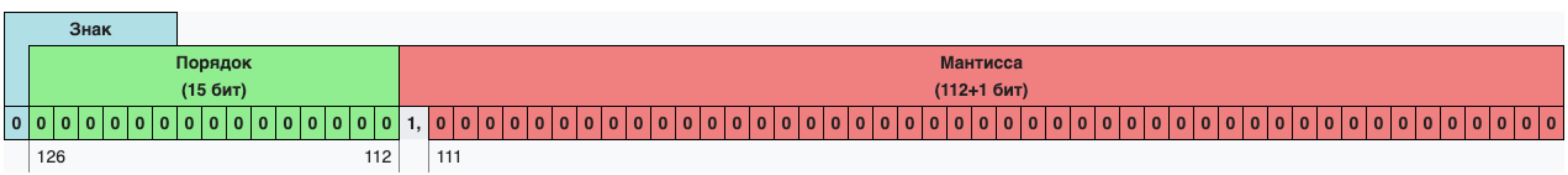
Число одинарной точности (Binary32, Single precision, float)



Число двойной точности (Binary64, Double precision, double)



Число четверной точности (Binary128, Quadruple precision)



ASCII

ASCII (American Standard Code for Informational Interchange — Американский стандартный код информационного обмена).

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

1963 год: 1 код = 7 бит. 2^7 = 128 символов.
Сейчас: 1 код = 8 бит = 1 байт. 2^8 = 256 символов.

ASCII Code Chart

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
0	NUL	SOH	STX	ETX	EOT	ENQ	ACK	BEL	BS	HT	LF	VT	FF	CR	SO	SI
1	DLE	DC1	DC2	DC3	DC4	NAK	SYN	ETB	CAN	EM	SUB	ESC	FS	GS	RS	US
2		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
3	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
4	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
5	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
6	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
7	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	DEL

HSE: ASCII**:** 48 53 45 или в двоичной: 010010000101001101000101

Десятичное число	*Восьмеричное число*	*Двоичная запись*	*Шестнадцатеричное число*	*Двоичная запись*
0	0	000	0	0000
1	1	001	1	0001
2	2	010	2	0010
3	3	011	3	0011
4	4	100	4	0100
5	5	101	5	0101
6	6	110	6	0110
7	7	111	7	0111
8	10	000	8	1000
9	11	001	9	1001
10	12	010	A	1010
11	13	011	B	1011
12	14	100	C	1100
13	15	101	D	1101
14	16	110	E	1110
15	17	111	F	1111

В настоящее время существует 5 кодовых таблиц для русских букв: Windows CP1251, MS – DOS CP866, КОИ – 8 (Код обмена информацией, 8-битный) (используется в OS UNIX), Mac (Macintosh), ISO (OS UNIX). Одним из первых стандартов кодирования кириллицы на компьютерах был стандарт **КОИ-8.**

Кодировка КОИ-8R

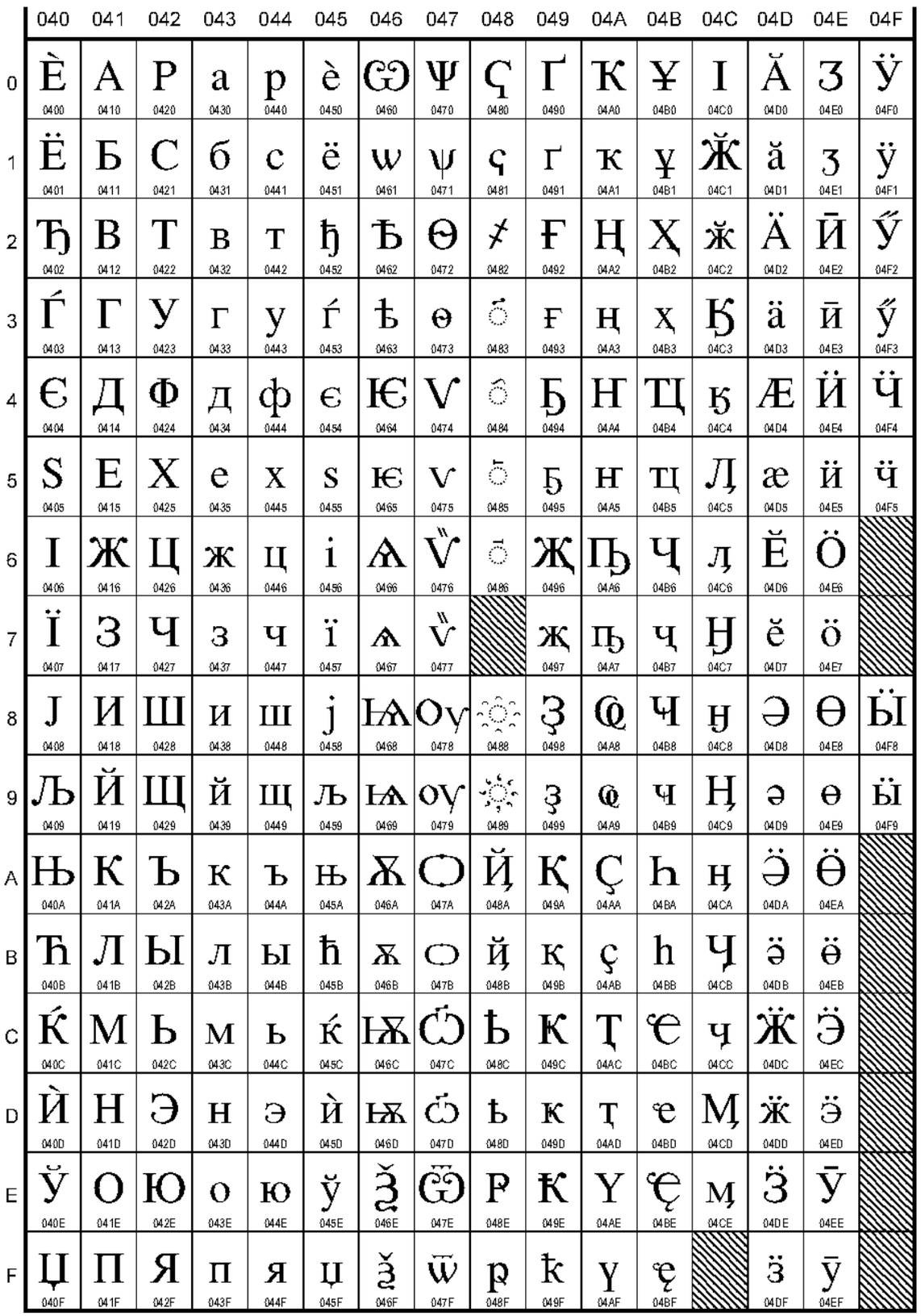
	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
80	2500 ─	2502 │	250C ┌	2510 ┐	2514 └	2518 ┘	251C ├	2524 ┤	252C ┬	2534 ┴	253C ┼	2580 ▀	2584 ▄	2588 █	258C ▌	2590 ▐
90	2591 ░	2592 ▒	2593 ▓	2320 ⌠	25A0 ■	2219 •	221A √	2248 ≈	2264 ≤	2265 ≥	A0	2321 ⌡	B0 °	B2 ²	B7 ·	F7 ÷
A0	2550 ═	2551 ║	2552 ╒	451 ё	2553 ╓	2554 ╔	2555 ╕	2556 ╖	2557 ╗	2558 ╘	2559 ╙	255A ╚	255B ╛	255C ╜	255D ╝	255E ╞
B0	255F ╟	2560 ╠	2561 ╡	401 Ё	2562 ╢	2563 ╣	2564 ╤	2565 ╥	2566 ╦	2567 ╧	2568 ╨	2569 ╩	256A ╪	256B ╫	256C ╬	A9 ©
C0	44E ю	430 а	431 б	446 ц	434 д	435 е	444 ф	433 г	445 х	438 и	439 й	43A к	43B л	43C м	43D н	43E о
D0	43F п	44F я	440 р	441 с	442 т	443 у	436 ж	432 в	44C ь	44B ы	437 з	448 ш	44D э	449 щ	447 ч	44A ъ
E0	42E Ю	410 А	411 Б	426 Ц	414 Д	415 Е	424 Ф	413 Г	425 Х	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О
F0	41F П	42F Я	420 Р	421 С	422 Т	423 У	416 Ж	412 В	42C Ь	42B Ы	417 З	428 Ш	42D Э	429 Щ	427 Ч	42A Ъ

Кодировка CP1251

	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
80	402 ђ	403 Ѓ	201A ‚	453 ѓ	201E „	2026 …	2020 †	2021 ‡	20AC €	2030 ‰	409 Љ	2039 ‹	40A Њ	40C Ќ	40B Ћ	40F Џ
90	452 ђ	2018 ‘	2019 ’	201C “	201D ”	2022 •	2013 –	2014 —	□	2122 ™	459 љ	203A ›	45A њ	45C ќ	45B ћ	45F џ
A0	A0	40E Ў	45E ў	408 Ј	A4 ¤	490 Ґ	A6 ¦	A7 §	401 Ё	A9 ©	404 Є	AB «	AC ¬	AD -	AE ®	407 Ї
B0	B0 °	B1 ±	406 І	456 і	491 ґ	B5 µ	B6 ¶	B7 ·	451 ё	2116 №	454 є	BB »	458 ј	405 Ѕ	455 ѕ	457 ї
C0	410 А	411 Б	412 В	413 Г	414 Д	415 Е	416 Ж	417 З	418 И	419 Й	41A К	41B Л	41C М	41D Н	41E О	41F П
D0	420 Р	421 С	422 Т	423 У	424 Ф	425 Х	426 Ц	427 Ч	428 Ш	429 Щ	42A Ъ	42B Ы	42C Ь	42D Э	42E Ю	42F Я
E0	430 а	431 б	432 в	433 г	434 д	435 е	436 ж	437 з	438 и	439 й	43A к	43B л	43C м	43D н	43E о	43F п
F0	440 р	441 с	442 т	443 у	444 ф	445 х	446 ц	447 ч	448 ш	449 щ	44A ъ	44B ы	44C ь	44D э	44E ю	44F я

Unicode

0400 **Cyrillic** **04FF**



- Результат сотрудничества Международной организации по стандартизации (ISO) с ведущими производителями компьютеров и программного обеспечения.
- В **Unicode** на кодирование символов отводится 16 бит, т.е. 2^16 = 65 536 кодов

Распространенные стандарты:

- **UTF-8**
- **UTF-16**
- **UTF-32**

One-hot. Унитарный код

Унитарный код - двоичный код фиксированной длины. Длина кода определяется количеством кодируемых объектов, то есть каждому объекту соответствует отдельный разряд кода, а значение кода положением 1 или 0 в кодовом слове.

Размер вектора = размеру словаря

Cat = [1,0,0,……,0]

Dog = [0,1,0,……..,0]

Bird = [0,0,0,……,1]

Label Encoding

Food Name	Categorical #	Calories
Apple	1	95
Chicken	2	231
Broccoli	3	50

→

One Hot Encoding

Apple	Chicken	Broccoli	Calories
1	0	0	95
0	1	0	231
0	0	1	50