Prof. Dr. Florian Künzner
Technical University of Applied Sciences Rosenheim, Computer Science

CA 2 – Data representation

The lecture is based on the work and the documents of Prof. Dr. Theodor Tempelmeier

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Goal

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Goal

CA::Data representation

- Important basics
- ASCII
- Unicode and UTF
- Data types: Numbers

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Important basics

Which numeral systems do you know?

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Important basics

Numeral systems

- DEC: 0, 1, ..., 9; e.g.: 291
- BIN: 0, 1; e.g.: 100100011
- HEX: 0, 1, ..., 9, A, B, ..., F; e.g.: 0x123

Conversion between:

- HEX <-> DEC
- **BIN <-> HEX**
- DEC <-> BIN

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Important basics - hints

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Important basics - short exercise 1/2

Convert `HEX:0xC0FE` to `BIN`.

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Important basics - short exercise 2/2

Convert `BIN:1100000011011110` to `HEX`.

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Binary system

Why is the binary (dual) system used in computer science?

Binary system for digits and characters

- Technically easy to realise (0/1)
- Well understood theoretical basis
  - Boolean algebra
  - Formal logic

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Subtraction is reduced to addition

Idea: Complementation and addition of the complement

Example: $11 - 6$ **in binary system**

```
11: -> 01011
 6: -> 00110
complement of 6:  11001
                +    1
                ------
                 11010
addition of 11 + (-6):
              11: 01011
              -6: 11010
              ---------
                 X00101 => 5
```

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Codes

Which codes for characters do you know?

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

ASCII (American Standard Code for Information Interchange)

Dec	Hx	Oct	Char	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr	Dec	Hx	Oct	Html	Chr
0	0	000	NUL (null)	32	20	040		Space	64	40	100	@	@	96	60	140	`	`
1	1	001	SOH (start of heading)	33	21	041	!	!	65	41	101	A	A	97	61	141	a	a
2	2	002	STX (start of text)	34	22	042	"	"	66	42	102	B	B	98	62	142	b	b
3	3	003	ETX (end of text)	35	23	043	#	#	67	43	103	C	C	99	63	143	c	c
4	4	004	EOT (end of transmission)	36	24	044	$	$	68	44	104	D	D	100	64	144	d	d
5	5	005	ENQ (enquiry)	37	25	045	%	%	69	45	105	E	E	101	65	145	e	e
6	6	006	ACK (acknowledge)	38	26	046	&	&	70	46	106	F	F	102	66	146	f	f
7	7	007	BEL (bell)	39	27	047	'	'	71	47	107	G	G	103	67	147	g	g
8	8	010	BS (backspace)	40	28	050	(	(	72	48	110	H	H	104	68	150	h	h
9	9	011	TAB (horizontal tab)	41	29	051	)	)	73	49	111	I	I	105	69	151	i	i
10	A	012	LF (NL line feed, new line)	42	2A	052	*	*	74	4A	112	J	J	106	6A	152	j	j
11	B	013	VT (vertical tab)	43	2B	053	+	+	75	4B	113	K	K	107	6B	153	k	k
12	C	014	FF (NP form feed, new page)	44	2C	054	,	,	76	4C	114	L	L	108	6C	154	l	l
13	D	015	CR (carriage return)	45	2D	055	-	-	77	4D	115	M	M	109	6D	155	m	m
14	E	016	SO (shift out)	46	2E	056	.	.	78	4E	116	N	N	110	6E	156	n	n
15	F	017	SI (shift in)	47	2F	057	/	/	79	4F	117	O	O	111	6F	157	o	o
16	10	020	DLE (data link escape)	48	30	060	0	0	80	50	120	P	P	112	70	160	p	p
17	11	021	DC1 (device control 1)	49	31	061	1	1	81	51	121	Q	Q	113	71	161	q	q
18	12	022	DC2 (device control 2)	50	32	062	2	2	82	52	122	R	R	114	72	162	r	r
19	13	023	DC3 (device control 3)	51	33	063	3	3	83	53	123	S	S	115	73	163	s	s
20	14	024	DC4 (device control 4)	52	34	064	4	4	84	54	124	T	T	116	74	164	t	t
21	15	025	NAK (negative acknowledge)	53	35	065	5	5	85	55	125	U	U	117	75	165	u	u
22	16	026	SYN (synchronous idle)	54	36	066	6	6	86	56	126	V	V	118	76	166	v	v
23	17	027	ETB (end of trans. block)	55	37	067	7	7	87	57	127	W	W	119	77	167	w	w
24	18	030	CAN (cancel)	56	38	070	8	8	88	58	130	X	X	120	78	170	x	x
25	19	031	EM (end of medium)	57	39	071	9	9	89	59	131	Y	Y	121	79	171	y	y
26	1A	032	SUB (substitute)	58	3A	072	:	:	90	5A	132	Z	Z	122	7A	172	z	z
27	1B	033	ESC (escape)	59	3B	073	;	;	91	5B	133	[	[	123	7B	173	{	{
28	1C	034	FS (file separator)	60	3C	074	<	<	92	5C	134	\	\	124	7C	174			|
29	1D	035	GS (group separator)	61	3D	075	=	=	93	5D	135	]	]	125	7D	175	}	}
30	1E	036	RS (record separator)	62	3E	076	>	>	94	5E	136	^	^	126	7E	176	~	~
31	1F	037	US (unit separator)	63	3F	077	?	?	95	5F	137	_	_	127	7F	177	&#127;	DEL

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Extended ASCII codes

128	Ç	144	É	160	á	176	▒	192	└	208	╨	224	α	240	≡
129	ü	145	æ	161	í	177	▒	193	┴	209	╤	225	ß	241	±
130	é	146	Æ	162	ó	178	▓	194	┬	210	╥	226	Γ	242	≥
131	â	147	ô	163	ú	179	│	195	├	211	╙	227	π	243	≤
132	ä	148	ö	164	ñ	180	┤	196	─	212	╘	228	Σ	244	⌠
133	à	149	ò	165	Ñ	181	╡	197	┼	213	╒	229	σ	245	⌡
134	å	150	û	166	ª	182	╢	198	╞	214	╓	230	µ	246	÷
135	ç	151	ù	167	º	183	╖	199	╟	215	╫	231	τ	247	≈
136	ê	152	ÿ	168	¿	184	╕	200	╚	216	╪	232	Φ	248	°
137	ë	153	Ö	169	⌐	185	╣	201	╔	217	┘	233	Θ	249	∙
138	è	154	Ü	170	¬	186	║	202	╩	218	┌	234	Ω	250	·
139	ï	155	¢	171	½	187	╗	203	╦	219	█	235	δ	251	√
140	î	156	£	172	¼	188	╝	204	╠	220	▄	236	∞	252	n
141	ì	157	¥	173	¡	189	╜	205	═	221	▌	237	φ	253	²
142	Ä	158	₧	174	«	190	╛	206	╬	222	▐	238	ε	254	■
143	Å	159	ƒ	175	»	191	┐	207	╧	223	▀	239	∩	255	

Source: www.LookupTables.com

[source: asciitable.com]

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

ASCII

ASCII - American Standard Code for Information Interchange

Any problems with ASCII?

Unicode

- International standard (ISO 10646)
- For every character one code
- In the long term: A digital code is defined for each meaningful character or text element of all known cultures, countries/languages, and character systems.
- Is constantly extended
- `http://www.unicode.org`

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Unicode

Character range:

first code U+00 0000

last code U+10 FFFF

Character sets

Name	Unit	Calculation	#chars	first	last
UCS-2	16 Bit	2^{16}	65536	U+0000	U+FFFF
UCS-4	17 Planes	$17 * 2^{16}$	1114112	U+00 0000	U+10 FFFF

Examples:

Unicode	Full number	Character
U+0041	00 0041	A
U+1F600	01 F600	😀

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Unicode 14.0 - Planes

Plane 0 00 0000-00 FFFF BMP Basic Multilingual Plane	Plane 1 01 0000-01 FFFF SMP Supplementary Multilingual Plane	Plane 2 02 0000-02 FFFF SIP Supplementary Ideographic Plane	Plane 3 03 0000-03 FFFF unassigned	Plane 4 04 0000-04 FFFF unassigned
Plane 5 05 0000-05 FFFF unassigned	Plane 6 06 0000-06 FFFF unassigned	Plane 7 07 0000-07 FFFF unassigned	Plane 8 08 0000-08 FFFF unassigned	Plane 9 09 0000-09 FFFF unassigned
Plane 10 0A 0000-0A FFFF unassigned	Plane 11 0B 0000-0B FFFF unassigned	Plane 12 0C 0000-0C FFFF unassigned	Plane 13 0D 0000-0D FFFF unassigned	Plane 14 0E 0000-0E FFFF SSP Supplementary Special-purpose Plane
Plane 15 0F 0000-0F FFFF SPUA-A Supplementary Private Use Area planes	Plane 16 10 0000-10 FFFF SPUA-A Supplementary Private Use Area planes			

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Unicode

Enter unicode characters

OS	Program	Keyboard shortcut
Linux	Terminal, xed, LibreOffice	`CTRL+SHIFT+U + HEX Number`
Windows	Microsoft Word, Excel, WordPad	`HEX Number + ALT+C`
macOS*	Console, Text	`ALT + HEX Number`

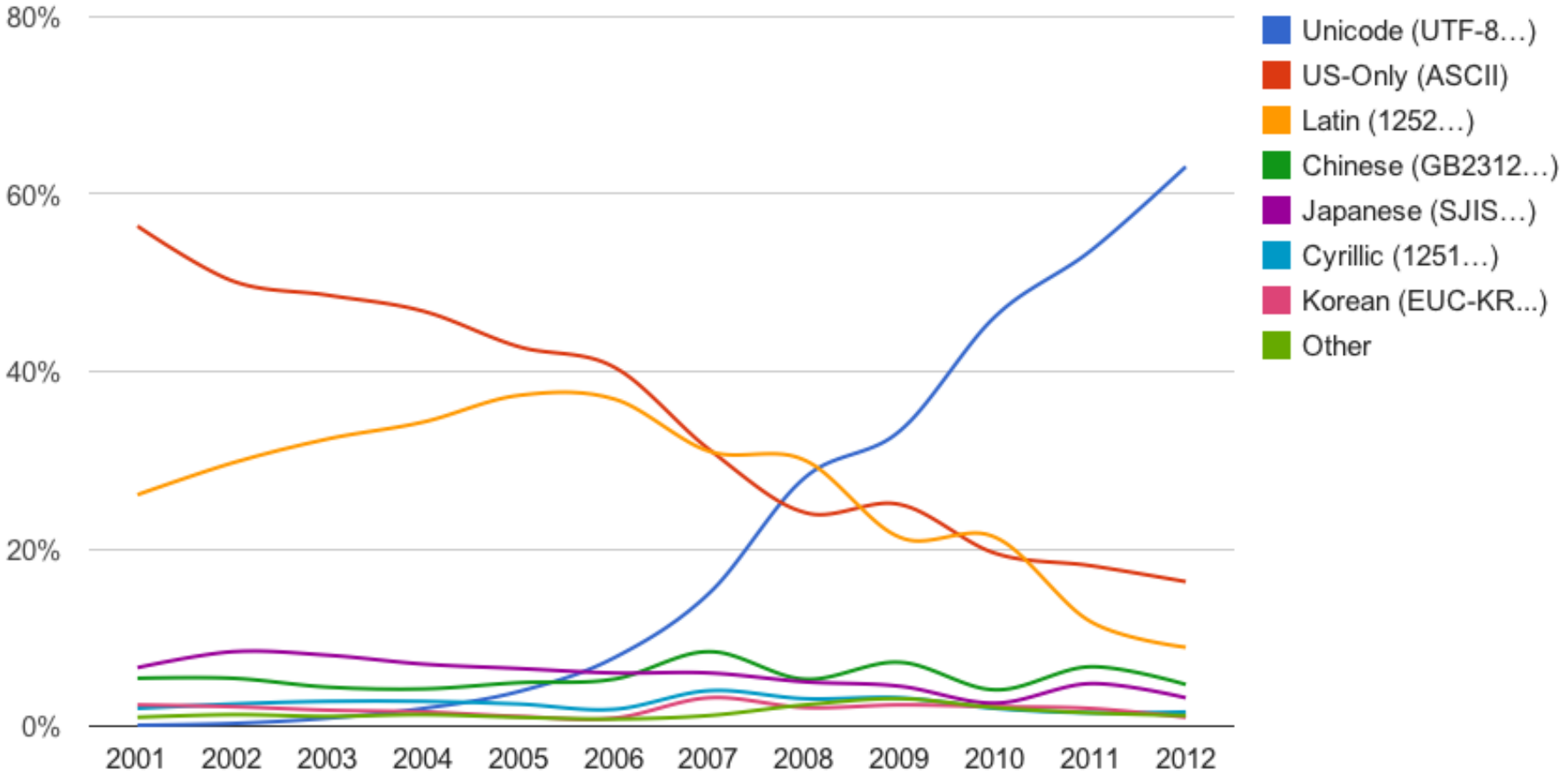
More shortcuts: wikipedia.org

*must be enabled as input source

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Unicode usage



[source: googleblog.blogspot.com], Link to current statistics: w3techs.com

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Unicode

Character set vs. character encoding?

Unicode vs **UTF**

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF - Unicode Transformation Format

UTF maps all unicode code points to a unique sequence of bytes.

Used for

- Store information into files, databases, ...
- Transfer data (websites, e-mail, ...)

Choice depends on

- Storage space
- Source code compatibility
- Interoperability with other systems
- Runtime for encoding/decoding

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF - Unicode Transformation Format

Overview of UTF encodings

Encoding	Bits	Length	Common use
UTF-8	8-bit	Variable length: 1 to 4 bytes	Internet, Linux
UTF-16	16-bit	Variable length: 2 or 4 bytes	Qt, Java, Tcl
UTF-32	32-bit	Fixed length: 4 bytes	

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-8

UTF-8 length

Number of bytes	Bits for code point	Unicode range	Comment
1	7	0 - 00 007F	Compatible with ASCII
2	11	80 - 00 07FF	
3	16	800 - 00 FFFF	
4	21	1 0000 - 10 FFFF	

UTF-8 encoding details

Unicode range	Byte 1	Byte 2	Byte 3	Byte 4
0 - 00 007F	0xxxxxxx			
80 - 00 07FF	110xxxxx	10xxxxxx		
800 - 00 FFFF	1110xxxx	10xxxxxx	10xxxxxx	
1 0000 - 10 FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-8 - example

Encode the „ü“ into UTF-8!

[ü: https://en.wikipedia.org/wiki/Latin-1_Supplement_(Unicode_block)]

```
1  ü -> 252 -> 0xFC
2
3  ü in Unicode:
4  U+00 00FC (8 bits -> 2 bytes required)
5  F    C
6  1111 1100
7
8  ü in UTF-8:
9  11000011 10111100
10 C   3    B   C     -> 0xC3BC
```

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-16

UTF-16 length

Number of bytes	Bits for code point	Unicode range	Comment
2	16	0 - 00 FFFF	
4	20	01 0000 - 10 FFFF	subtraction required: U+XXXXXX - 0x10000

UTF-16 encoding details

Unicode range	Byte 1	Byte 2	Byte 3	Byte 4
0 - 00 FFFF	xxxxxxxx	xxxxxxxx		
	High surrogate		Low surrogate	
01 0000 - 10 FFFF	110110xx	xxxxxxxx	110111xx	xxxxxxxx

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-16 - example

Encode the „😀“ (U+1F600) into UTF-16!

```
4 byte variant and therefore correction required:
0x1F600 - 0x10000 = 0xF600

F    6    0    0
1111 0110 0000 0000

In UTF-16:
 High surrogate      Low surrogate
11011000 00111101  11011110 00000000
D   8    3   D     D   E    0   0     -> 0xD83DDE00
```

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-32

UTF-32 length

Number of bytes	Bits for code point	Unicode range	Comment
4	21	00 0000 - 10 FFFF	directly representable

UTF-32 encoding details

Unicode range	Byte 1	Byte 2	Byte 3	Byte 4
0 - 10 FFFF	00000000	000xxxxx	xxxxxxxx	xxxxxxxx

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

UTF-32 - example

Encode the „😀“ (U+1F600) into UTF-32!

```
1 Only the 4 byte variant exists
2 0x1F600
3
4 1    F    6    0    0
5 0001 1111 0110 0000 0000
6
7 In UTF-32:
8 00000000 00000001 11110110 00000000
9 0    0    0    1    F    6    0    0     -> 0x0001F600
```

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Numbers

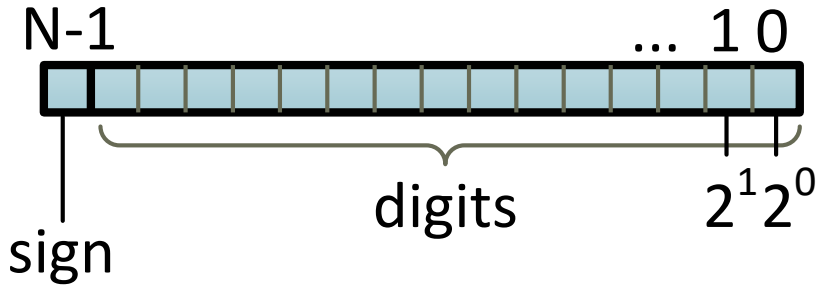
Type	Common data type	Realisation
Integer	unsigned int, int, ...	Hardware: ALU
Floating point – binary	float, double, ...	Hardware: FPU
Floating point – decimal	decimal32, decimal64, ...	Mostly in software
Fixed point – binary	Often not well integrated	Mostly in software
Fixed point – decimal	Often not well integrated	Mostly in software

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Integer (signed)

Example: `short int`



Positive number: The weight for position i is 2^i
Negative number: The sign is interpreted as -2^N
Example `short int`: Minimum: -32768; Maximum: 32767

limits: `http://www.cplusplus.com/reference/climits`

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Integers

Any problems with integers?

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Fixed width integer types (since C99)

Available types:

Bits	signed	unsigned
8	int8_t	uint8_t
16	int16_t	uint16_t
32	int32_t	uint32_t
64	int64_t	uint64_t

Example:

```
1 #include <stdint.h>
2 ...
3 int16_t val_16 = 5;
```

More details: https://en.cppreference.com/w/cpp/types/integer

Floating point – binary

Usually scientific numbers with mantissa and exponent.
Requires hardware support (FPU - floating point unit).

Format: $x = m \cdot B^e$ (m = mantissa, B = basis, and e = exponent)

Examples:

- C: `float x;`
- Ada: `x: float`

Floating point – binary

Floating point binary formats are defined in the **IEEE Standard for Floating-Point Arithmetic (IEEE 754)**.

Name	Common name	Number of bits	Characteristic	Mantissa
binary16	Half precision	16	5 bits; $c = e + 15$	10 bits
binary32	Single precision	32	8 bits; $c = e + 127$	23 bits
binary64	Double precision	64	11 bits; $c = e + 1023$	52 bits
binary128	Quadruple precision	128	15 bits; $c = e + 16383$	112 bits
binary256	Octuple precision	256	19 bits; $c = e + 262143$	236 bits

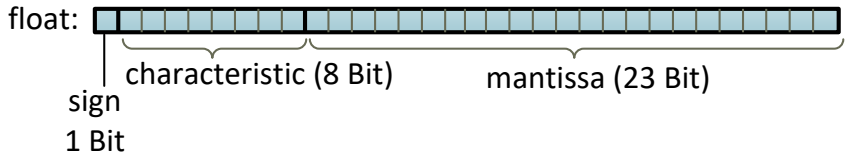
IEEE 754 on Wikipedia: `https://en.wikipedia.org/wiki/IEEE_754`

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Floating point – binary

Example: `float (single precision)`



Exponent	$-126, \ldots, +127$	Exponent is represented via the characteristic
Characteristic	$c = e + 127$	
Mantissa	$1 \leq m < B$	Is normalised in the binary system: `1.MMM...M` Advantage: `1` doesn't have to be saved!

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Floating point – binary

Convert the decimal number 1.75 into the binary32 (float) representation.

```
1.75 -> binary:
01.11000...0   -> it has already the required form
                  of 1.MMM...M (=> e=0)

c = e + 127 = 0 + 127 = 127

S|C       |M
0|01111111|11000000000000000000000|

Hex representation:
0x3fe00000
```

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Floating point – binary

Let's do some (binary) floating point number crunching.

Nr.	Code	different	equal
1	`36.2 != 36.2`		
2	`0.362 * 100.0 != 36.2`		
3	`0.362 * (100.0 / 100.0) != 0.362`		
4	`(0.362 * 100.0) / 100.0 != 0.362`		

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Floating point – decimal

Floating point decimal formats are defined in the **IEEE Standard for Floating-Point Arithmetic (IEEE 754)**.

Format: $x = (-1)^{\text{signbit}} \times 10^{\text{exponentbits}_2 - 101_{10}} \times \text{truesignificand}_{10}$

Name	Number of decimal digits	Exponent min.	Exponent max.
decimal32	7	-95	+96
decimal64	16	-383	+384
decimal128	34	-6143	+6144

IEEE 754 on Wikipedia: `https://en.wikipedia.org/wiki/IEEE_754`

- Possible in gnu C with `_Decimal32`, `_Decimal64`, and `_Decimal128`
- Example C: `_Decimal32 x = 0.1df;`
- Possible in gnu C++ with `decimal32`, `decimal64`, and `decimal128`
- Example C++: `std::decimal::decimal32 x(0.1);`

More details on the format (on Wikipedia): `https://en.wikipedia.org/wiki/Decimal32_floating-point_format`

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Floating point – decimal

Let's do some (decimal) floating point number crunching.

Nr.	Code	different	equal
1	`36.2 != 36.2`		
2	`0.362 * 100.0 != 36.2`		
3	`0.362 * (100.0 / 100.0) != 0.362`		
4	`(0.362 * 100.0) / 100.0 != 0.362`		

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Fixed point

Fixed point numbers have a fixed imaginary point that is not moved.

Usage:

- Areas where rounding errors must be avoided (e.g. commercial applications)
- If no floating point hardware (FPU) is available (e.g. in embedded systems)
- Devices use the numbers in this format anyway (e.g. analog/digital converter)

Two variants:

Type	Usage
Binary fixed point	technical
Decimal fixed point	economical

CAMPUS Rosenheim
Computer Science

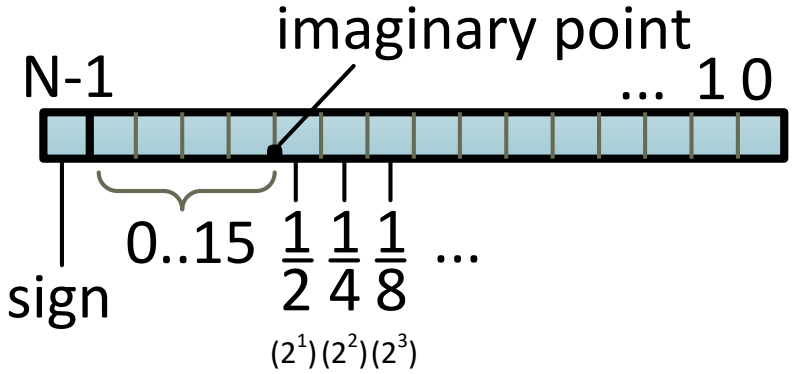
Technische Hochschule Rosenheim
Technical University of Applied Sciences

Fixed point – binary

Uses integers with an imaginary binary point.

Often without specialised hardware: *Poor man's floating point*.

Ada: `type analog_input is delta 0.125 range -16.0..15.0;`



[C++ library example: Compositional Numeric Library]

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

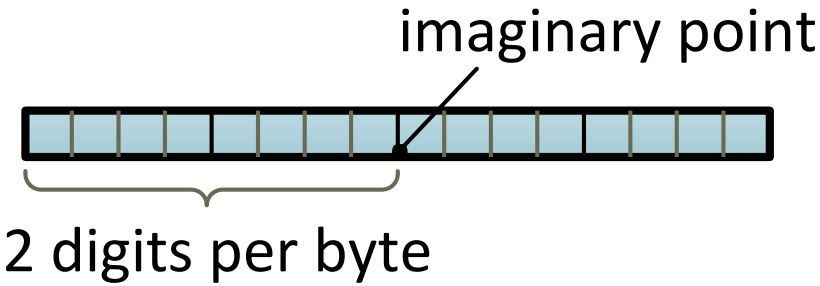
Fixed point – decimal

Uses the binary coded decimal (BCD) system with an imaginary decimal point and BCD arithmetic.

Used in IBM main frame. Sometimes there exists specialised hardware.

BCD: Every digit $(0-9)$ is represented by 4 bits

Ada: `type money is delta 0.01 digits 8;`



[C++ library example: Decimal data type for C++]

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

When to use what?
A first try for a property overview

CAMPUS Rosenheim
Computer Science

Technische Hochschule Rosenheim
Technical University of Applied Sciences

Summary and outlook

Summary

- Important basics
- ASCII
- Unicode and UTF
- Data types: Numbers

Outlook

- Logical hardware