

# MIE-ARI

## (Computer Arithmetic – Homework 5)

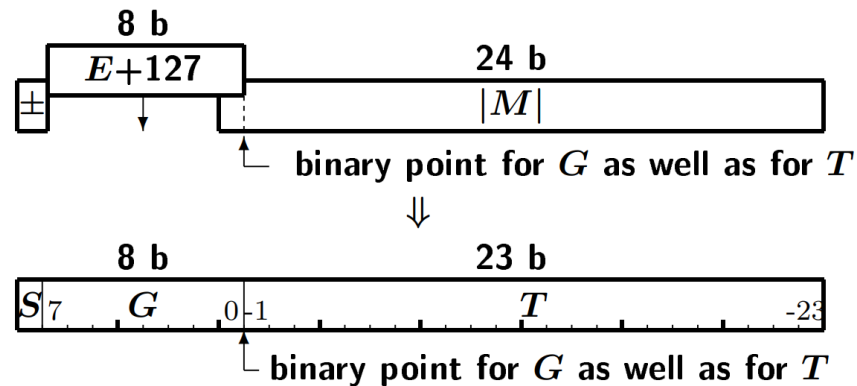
### Floating point

Pavel Kubalík  
Department of Digital Design  
Faculty of Information Technology  
Czech Technical University in Prague

<https://courses.fit.cvut.cz/MIE-ARI/>

# Task 1 – IEEE Std 754-2008 - Example

Calculate image of 32-bits number represented in IEEE Std 754 format.



	$A$
$G = 0 \dots 0_2$	$(-1)^S \cdot T \cdot 2^{-K+1}$
$G = 1 \dots 1_2$ a $T = 0$	$(-1)^S \cdot \infty$
$G = 1 \dots 1_2$ a $T \neq 0$	<b>NaN</b>
else (viz FP - 14)	$(-1)^S \cdot (1 + T) \cdot 2^{G-K}$

a) 0000 0000 0100 0..0

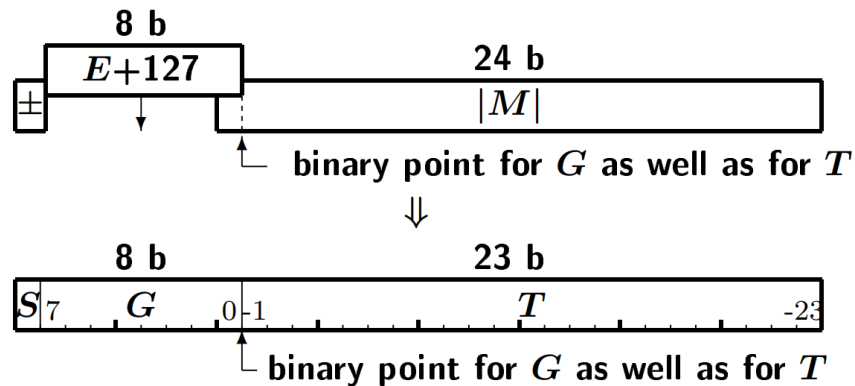
b) 0000 0000 1100 0..0

c) 1100 1000 0100 0..0

Advice: Use the information in lecture 6 (Floating point), in slides 14-16.

# Task 2 –IEEE Std 754-2008

Calculate image of 32-bits number represented in IEEE Std 754 format.



	$A$
$G = 0 \dots 0_2$	$(-1)^S \cdot T \cdot 2^{-K+1}$
$G = 1 \dots 1_2$ a $T = 0$	$(-1)^S \cdot \infty$
$G = 1 \dots 1_2$ a $T \neq 0$	<b>NaN</b>
else (viz FP – 14)	$(-1)^S \cdot (1 + T) \cdot 2^{G-K}$

a) 0000 0000 0110 0..0

b) 0100 0000 1100 0..0

c) 1100 0001 0101 0..0

Advice: Use the information in lecture 6 (Floating point), in slides 14-16.

# Task 2 – Addition - Example

- Convert two decimal numbers to binary system.
- Add them using GRS bits.
- Round the result to prefer greater value.
- Round the result to prefer even value.

$$91,5_{10} + 114,25_{10} =$$

ABS(M)										G	R	S	E
1	,	0	1	1	0	1	1	1	0	0	0	0	6
1	,	1	1	0	0	1	0	0	1	0	0	6	

Advice: Use the information in lecture 6 - Floating point.

## Task 3 – Addition

- Convert two decimal numbers to binary system.
- Add them using GRS bits.
- Round the result to prefer greater value.
- Round the result to prefer even value.

$$45,75_{10} + 3,5625_{10} =$$

	ABS(M)	G	R	S	E
,					
,					

Advice: Use the information in lecture 6 - Floating point.

## Task 4 – Addition

- Convert two binary numbers to decimal system.
- Add them using GRS bits.
- Round the result to prefer greater value.
- Round the result to prefer even value.

ABS(M)										G	R	S	E
1	,	0	1	1	0	1	1	1	0	0	0	0	8
1	,	1	1	0	0	1	0	0	1	0	0	0	4

Advice: Use the information in lecture 6 - Floating point.

# Notes

# Notes