NATIONAL RESEARCH UNIVERSITY

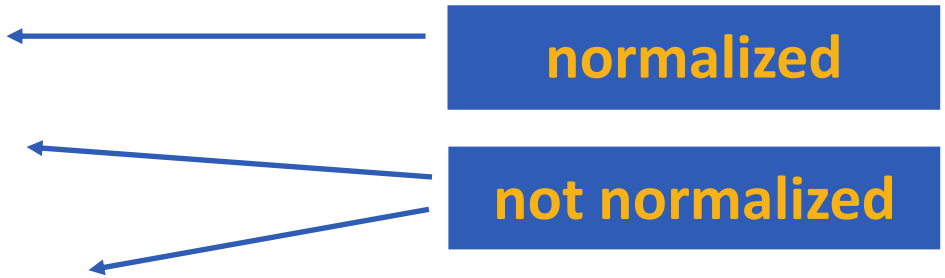
Computer Architecture and Operating Systems
Lecture 9: Floating-Point Format

Andrei Tatarnikov

atatarnikov@hse.ru
@andrewt0301

Floating-Point Format

- Representation for non-integral numbers
  - Including **very small** and **very large** numbers
- Like scientific notation
  - -2.34×10^{56}
  - $+0.002 \times 10^{-4}$
  - $+987.02 \times 10^{9}$



- In binary
  - $\pm 1.xxxxxxx_2 \times 2^{yyyy}$
- Types **`float`** and **`double`** in C

Floating-Point Standard

- Defined by **IEEE Std 754-1985**
- Developed in response to divergence of representations
  - **Portability** issues for scientific code
- Now almost universally adopted
- Two representations
  - **Single precision** (32-bit)
  - **Double precision** (64-bit)

Any Questions?

```
            .text
__start:    addi t1, zero, 0x18
            addi t2, zero, 0x21
cycle:      beq t1, t2, done
            slt t0, t1, t2
            bne t0, zero, if_less
            nop
            sub t1, t1, t2
            j cycle
            nop
if_less:    sub t2, t2, t1
            j cycle
done:       add t3, t1, zero
```