Floating point numbers representation

IEEE754

IEEE754

Single-Precision (32 бит); B = 127		
S	E (8 bit)	M (23 bit)
Double-Precision (64 бит); B = 1023		
S	E (11 bit)	M (52 bit)

$$Value = (-1)^{S} \cdot 2^{E-B} \cdot (1 + M / 2^{52})$$

Special values

S	E	M	Value
0	0	0	+0
1	0	0	-0
0	11…11	0	+ ∞
1	11…11	0	- ∞
0	11…11	≠ 0	Signaling NaN
1	11…11	≠ 0	Quiet NaN
0	0	≠ 0	Denormalized values
1	0	≠ 0	

Denormalized values

Single-Precision (32 бит)		
S	0000 0000	M (23 bit)
Double-Precision (64 бит)		
S	000 0000 0000	M (52 bit)

$$\text{Value} = (-1)^S \cdot M / 2^{52}$$

Operations: multiplication

$\langle S,E,M \rangle = \langle S_1,E_1,M_1 \rangle \cdot \langle S_2,E_2,M_2 \rangle$

1. Calculate

   $S = S_1 \text{ ^ } S_2$

   $E = E_1 + E_2$

   $M = 1.M_1 \cdot 1.M_2$

2. { M >>= 1; E++ } while M overflows

Operations: addition

$<S,E,M> = <S_1,E_1,M_1> \cdot <S_2,E_2,M_2>$

1. Calculate

   E_{diff}=E1 - E2
2. Normalize M_2 to E_{diff} bits
3. Values:

   $E = E_1$

   $M = M_1$ +/- M_2

   $S = sign(-1^{S1}M_1 + -1^{S2}M_2)$

FPU implementations

- Extended command set (ARM VFP):
  - Additional commands
  - Additional 32 registers
- Coprocessor (x86): (gcc -mfpmath=387)
  - Commands that CPU gives to FPU
  - Interaction through stack
- SSE commands (Pentium-III+, x86-64): (gcc -msse -mfpmath=sse)
  - xmm registers used
  - Scalar commands SSE used

Precision

- Floating point
  - Single precision – 32 bit ≈ 10^{37}
  - Double Precision - 64bit ≈ 10^{307}
- Fixed point