2.0 Floating point numbers

$$y = M \times x^{E} \quad E \leftarrow \text{exponent.} \in \mathbb{Z}$$

M ← mantissa (fixed point)

x ← base

1.38×10^{-4}

2.71×10^{3}

1.38E-4

2.71E3

$M \equiv Q0.m$ signed

$E \equiv Qn.0$ signed

Example

$1.011e011 = (-1 + 0.25 + 0.125) \times 2^{3}$

$= -0.625 \times 2^{3}$

$= -5.00_{10}$

$0.101e011$

$0,101_{2} \times 2_{10}^{3_{10}} = 01.01_{2} \times 2_{10}^{2_{10}} = 010.10_{2} \times 2_{10}^{1_{10}} =$

$= 0101$

$1.50 \times 10^{3} \Rightarrow 1.50 \times 10^{3}$

$+ 1.01 \times 10^{2} \quad + 0.10 \times 10^{3}$

$= 15.0 \times 10^{2}$

$+ 1.01 \times 10^{2}$

$0.733 \times 10^{5} = 7.330 \times 10^{4}$

$+ 0.015 \times 10^{7} = 1.5 \times 10^{5}$

Binary examples

Normalising:

$111101e00100 = (-1 + 0.5 + 0.25 + 0.125 + 0.03125)$

$= 111010e00011 \quad \times 2^4 = -1.5_{10}$

$= 111010e00010$

$= 101000e00001 = (-1 + 0.25) \times 2^1 = -0.75 \times 2 = -1.5$

2.1 Normalisation

Act of maximising precision.

left shift mantissa and subtract exponent.
UNTIL MSB = (MSB_1).

2.2 Justification

Act of matching exponents

Arithmetic right shift and increase exponent
UNTIL exponent matches target exponent

Justification example

$1011011\,00 \epsilon 0000$ to $\epsilon 0111$

$= 1101101\,10 \epsilon 0001$

$= 1110110\,11 \epsilon 0010$

$= 1111011\,01 \epsilon 0011$

$= 1111101\,11 \epsilon 0100$

$= 1111110\,11 \epsilon 0101$

$= 1111111\,01 \epsilon 0110$

$= 11111110 \epsilon 0111$

$00010\,1 \epsilon 0010 + 01100\,1 \epsilon 0011$

$0.101\,00 \epsilon 0000$
$+ 0.110\,01 \epsilon 0011$

$= 0.000\,10 \epsilon 0011$
$+ 0.110\,01 \epsilon 0011$
$0.111\,01 \epsilon 0011$

$0.101 \epsilon 000$
$+ 0.101 \epsilon 000$
overflow : $01 \epsilon 000$
$0.101 \epsilon 001$

Overflow example

$0.101 \epsilon 000$
$+ 0.101 \epsilon 000$

2.3 Floating point addition and subtraction

1. Normalise.
2. Justify on max exponent.
3. Add/Subtract.
4. Deal with carry/overflow.
5. Normalise answer.

2.4 FP Multiplication

$$y_1 = M_1 \times x^{E_1}$$
$$y_2 = M_2 \times x^{E_2}$$
$$z = y_1 \times y_2 = (M_1 \times M_2) \times x^{E_1 + E_2}$$

1. Normalise
2. Fixed point multiplication of mantissas.
3. Fixed point addition of exponents.
4. Normalise

Example

101 e 0110

111 e 0100 = 100 e 0010

$M_1 \times M_2$

Q0.2 011

Q1.2 x 0100

000

0000

Σ 01100

000000

06.1100 Q1.4

0.1100 Q0.4

Add exponents

0 0110

0 0010

0 1000 overflow, adding bit to preserve sign

= 01000

z = 01100 e 01000