Lecture Notes 9/4 floating point conversions

Wednesday, September 4, 2024 18:06

$$\begin{array}{r} \overset{1}{1}\overset{1}{2}3_{10} \\ +\ 789_{10} \\ \hline 912 \end{array} \qquad \begin{array}{r} \overset{1}{1}\overset{1}{2}3_{8} \\ +\ 567_{8} \\ \hline 712_{8} \end{array} \qquad \begin{array}{r} \overset{1}{1}\overset{1}{2}3_{16} \\ +\ DEF_{16} \\ \hline F12 \end{array} \qquad \begin{array}{r} \overset{1}{1}0\overset{1}{1}\overset{1}{1} \\ +\ 1001 \\ \hline 10100 \end{array}$$

$5_{10} \rightarrow 101_2$ $\qquad$ $101_2 \times 2^1$

$10_{10} \leftarrow 1010_2$ shifted 1 place to right

Complements

74_{10}

$$\begin{array}{r} 999 \\ -\ 74 \\ \hline \underline{9}25 \\ +\ 1 \\ \hline \underline{9}26 \end{array}$$

9's complement

10's complement

$\rightarrow -74_{10}$

Prove by adding

$$\begin{array}{r} \overset{1}{}\overset{1}{7}4 \\ +\ \underline{9}26 \\ \hline \underline{0}00 \end{array}$$

$\underline{9}26$

$$\begin{array}{r} \cdots\cdots 99999926 \\ 74 \\ \hline \cdots\cdots 000000 \end{array}$$

$74_8 \xrightarrow{?} -74_8$.

$$\begin{array}{r} \underline{7}77 \\ -\ 74 \\ \hline \underline{7}03 \\ +\ 1 \\ \hline \underline{7}04 \end{array}$$

7's comp

8's comp

Prove by adding

$$\begin{array}{r} \overset{1}{1}74_8 \\ +\ \underline{7}04_8 \\ \hline \cdots 000 \end{array}$$

74_{16}

$$\begin{array}{r} \underline{F}FF \\ -\ 74 \\ \hline \underline{F}8B \\ +\ 1 \end{array}$$

15's comp

1111
0100

1011

$$\begin{array}{r} - \quad 74 \\ \hline F\,8\,B \end{array}$$ 15's comp

$$\begin{array}{r} +1 \\ \hline F\,8\,C \end{array}$$ 16's comp

Prove →

$$\begin{array}{r} 74 \\ + \; F8C \\ \hline \ldots 000 \end{array}$$

101_2

$$\begin{array}{r} 1111 \\ - \; 101 \\ \hline 1010 \end{array}$$ 1's comp

$$\begin{array}{r} 1 \\ \hline 1011 \end{array}$$ 2's comp

0000 0101

1111 1010 1's comp

+1

1111 1011

Prove →

$$\begin{array}{r} 11111011 \\ + \; 101 \\ \hline \ldots 00000000 \end{array}$$

Convert floats

7.1_{10}

int protion

$7_{10} = 7_{16}$

$7_{10} = 7_8$

$7_{10} = 111_2$

decimal

if base 16

$.1_{10} \rightarrow .1 \times 16 = 1.6$

$.6 \times 16 = 9.6$

$.6 \times 16 = 9.6$

$.1_{10} = .19999999 \ldots$

$= .1\underline{9}_{16}$

$.1_{10} = .1999$

$1/16 + 9/16^2 + 9/16^3 + 9/16^4 \ldots$

$1/16 + 9/16^2 (1 + 1/16 + 1/16^2 + \ldots)$

$\sum_{i=0}^{\infty} c^i \Rightarrow c = 1/16$

$\frac{1}{1-c}$

$= \frac{1}{16} + \frac{9}{16^2}\left(\frac{1}{1-1/16}\right) \left(\frac{1}{15/16}\right) = 16/15$

Proof → $1/16 + 9/16^2 \; 16/15 = 1/16 + 9/240 = \frac{15+9}{240} = \frac{24}{240} = 1/10 = .1$

$.1_{10} = .1\underline{9}$

$.19_{16} < .19_{16} < .1A_{16}$

$.2_{10} \xrightarrow{16}$ $.2 \times 16 = 3.\underline{2}$

$.2 \times 16 = 3.\underline{2}$

So

$.2_{10} = .3_{16}$

$= 3/16 + 3/16^2$

Sci note

$.2_{10} = .3_{16} = .3\ 3\ 3\ 3\ 3$

$.0011\ 0011\ 0011\ 0011\ 0011$

$.\underline{146}\ 3\ 1\ 4\ 6\ 3$

$.\underline{3}_{16}$

$.\underline{0011}_2$

$.\underline{1463}_8$

$.3_{16} = .3^0 \times 16^0 = 3.3 \times 16^{-1} = 33.3 \times 16^{-2}$

$= .3333 \times 16^4$

IEEE 754 format

4 byte float

"Power"

characteristic

mantissa

sign

$\_\ [\_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ |\ \_]\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ |\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ |\ \_\ \_\ \_\ \_\ \_\ \_\ \_\ \_$

+127

shifted by 127

127-4 $\quad 0\ 1\ 1\ 1\ 1\ 1\ 1\ 1$

$-\quad\quad\quad\quad\quad 1\ 0\ 0$

$.1_{10} = .19_{16} = .0001\ 1001\ 1001\ 1001\ 1001$

$1.1001100110 01 \times 2^{-4}$

Round up

$0\ [0\ 1\ 1\ 1\ 1\ 0\ 1\ |\ 1]\ .1\ 0\ 0\ 1\ 1\ 0\ 0\ |\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 0\ |\ 1\ 1\ 0\ 0\ 1\ 1\ 0\ 1$

0 [0 1 1 1 1 1 0 1 | 1] .1 0 0 1 1 0 0 | 1 1 0 0 1 1 0 0 | 1 1 0 0 1 1 0 1

3 D C C C C C D

$.2 = .3333 = .0011\,0011\,0011\,0011$

$1.10011001100 11 \times 2^{-3}$

$$\begin{array}{rr} 01111111 & 127 \\ -\quad 11 & -3 \\ \hline 01111100 & \end{array}$$

round up

0 [0 1 1 1 1 1 0 | 0] 1 0 0 1 1 0 0 | 1 1 0 0 1 1 0 0 | 1 1 0 0 1 1 $\boxed{1\ 0}$

3 E 4 C C C C D

$7.2_{10} = 7.3_{16} = 0111.0011\,0011$

1.1100110011×2^2

127+2

no round

0 [1 1 1 1 1 1 1 | 1] 1 1 0 0 1 1 0 | 0 1 1 0 0 1 1 0 | 0 1 1 0 0 1 1 $\boxed{0}$

1 0

0 [1 0 0 0 0 0 0 | 1]

4 0 E 6 6 6 6 6