Converting ints to floats

CS520

Dept. of Computer Science
Univ. of New Hampshire

<u>i2f</u> <u>example</u> 0x3456789A → Float

$$\boxed{0}011\ 0100\ 0101\ 0110\ 0111\ 1000\ 1001\ 1010.\times 2^{0}$$

→

$$1.\underline{10100\ 0101\ 0110\ 0111\ 1000\ 10}\,\big|\,\underline{01}\ \underset{2}{\underline{1010}}\times 2^{29}$$

~~$\times 2^{29}$~~

in this case truncate extra bits

guard bits

summarized as either
∅ or 1
∅ – all other bits are ∅
1 – otherwise

$$\text{actual exp} = \text{stored exp} - 127$$

$$29 = \text{stored exp} - 127$$

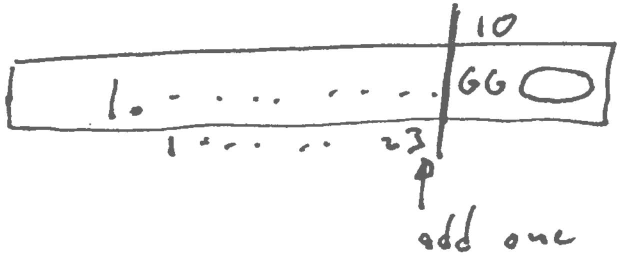
$$\text{stored exp} = 156_{10} = 10011100_{2}$$

$$0100\,|\,1110\,|\,0101\,|\,0001\,|\,0101\,|\,1001\,|\,1110\,|\,0010$$

$$4\quad E\quad 5\quad 1\quad 5\quad 9\quad E\quad 2$$

IEEE Floating-point rounding

round to even



guard bits*	sticky bit**	action
00	0 or 1	truncate
01	0 or 1	truncate
10	0	round to even
10	1	add one
11	0 or 1	add one

* two highest bits to be discarded

** summarizes all other discarded bits:
if all 0's then sticky bit is 0.
else sticky bit is 1

truncate – just discard the bits

round to even – if low bit in significand
is 1, add 1
else do nothing

add one – add one to the significand

note: could be carry out the top requiring re-normalization

shift right & adjust the exp.

$$\begin{array}{r} 1.1 \longrightarrow 1 \\ +1 \\ \hline 10.0 \longrightarrow 0 \\ \longrightarrow \end{array}$$

work these cases yourself

0x345678A0 → 4E5159E2

guard bits : 10
sticky bit : 0 } round to even, but low bit is Ø (i.e. already even) so nothing done

0x345678EØ → 4E5159E4

$$\left.\begin{array}{ll}\text{guard bits:} & 10\\ \text{sticky bit:} & 0\end{array}\right\}$$

round to even, low bit is 1 (i.e. odd), so add 1 to that position

0x345678A1 → 4E5159E3

$$\left.\begin{array}{l}\text{guard bits : 10}\\ \text{sticky bit : 1}\end{array}\right\}\ \begin{array}{c}\text{add one to low bit}\\ \text{position}\end{array}$$

0x345678B0 → 4E5159E3

$$\left.\begin{array}{l}\text{guard bits : 11}\\ \text{sticky bit : }\emptyset\end{array}\right\}\ \begin{array}{c}\text{add one to low}\\ \text{bit position}\end{array}$$