Converting ints to floats

CS520
Dept. of Computer Science
Univ. of New Hampshire
If example $0x3456789A \rightarrow \text{Float}$

$$0 \downarrow 1010001010101111100010011010 \times 2^0$$

$$1.1010001010111100010011010 \times 2^{29}$$

In this case truncate extra bits

Guard bits

summarized as either

$\phi$ or $1$

if then

all then

6 bits or

$\phi$

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

$29 = \text{stored } \exp$

$\text{stored } \exp = 156_{10} = 10011100_2$

$0100111001010010110010111100010$

$4 E 5159E2$
IEEE Floating-Point Rounding

**Round to even**

<table>
<thead>
<tr>
<th>Guard bits*</th>
<th>Sticky bit**</th>
<th>Action</th>
</tr>
</thead>
<tbody>
<tr>
<td>00</td>
<td>0 or 1</td>
<td>truncate</td>
</tr>
<tr>
<td>01</td>
<td>0 or 1</td>
<td>truncate</td>
</tr>
<tr>
<td>10</td>
<td>0</td>
<td>round to even</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>add one</td>
</tr>
<tr>
<td>11</td>
<td>0 or 1</td>
<td>add one</td>
</tr>
</tbody>
</table>

* 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

1.1 \rightarrow (\text{shift & adjust right} & \text{the exp.})

10.0 \rightarrow 0
Work these cases yourself

\[ 0 \times 345678AO \rightarrow 4E5159E2 \]

Guard bits: 10
Sticky bit: 0

Round to even, but low bit is 0 (i.e. already even) so nothing done
$0x345678E0 \rightarrow 4E5159E4$

Guard bits: 10
Sticky bit: 0

Round to even, low bit is 1 (i.e. odd), so add 1 to that position.
\[ 0x345678A1 \rightarrow 4E5159E3 \]

\[
\text{guard bits: 10} \]
\[
\text{sticky bit: 1} \]

\[
\text{add one to low bit position}
\]
0x345678B0 $\rightarrow$ 4E5159E3

guard bits: 11
sticky bit: 0
add one to low bit position