

Rounding

- Floating Point Number are normally approximations
 - Why?
 - Infinite variety of real numbers, but
 - 2^{53} ways of expression in double precision fp
 - IEEE 754 Rounding Modes of Approximation
 - 2 extra bits on the right during intermediate
 - guard bit and round bit

- Guard and Round bits - Illustration

- Addition Example (3 significant decimal digits) $2.56 \times 10^0 + 2.34 \times 10^2$

- Without Guard and Round

$$\begin{array}{r} 0.02 \\ 2.34 \\ \hline 2.36 \end{array} \rightarrow 2.36 \times 10^2$$

- With Guard and Round

$$\begin{array}{r} \text{guard} \\ \downarrow \\ 0.02\textcolor{blue}{5}\textcolor{red}{6} \\ 2.34\textcolor{blue}{0}\textcolor{red}{0} \\ \hline 2.36\textcolor{blue}{5}\textcolor{red}{6} \end{array} \rightarrow 2.37 \times 10^2$$

guard
 round

00-49 : round down
 51-99 : round up

Accuracy in floating points

- Measure of accuracy
 - the number of errors in the LSBs of the significands
 - “units in the last place” ----> ulp
 - Problem when a number is half-way in-between (i.e., 0.5---->0? 1?)
 - Norm --round to nearest even number
 - a third bit - “sticky” bit (next to the *guard* and *round*)
 - the sticky bit is set whenever there are nonzero bits to the right of the round bit
- Sticky bit example
$$5.01 \times 10^{-1} + 2.34 \times 10^{-2} \text{ (3 significant decimal digits)}$$
- Without sticky bit

0.00501	---->	0.0050
2.3400		2.3400
		2.3450 ----> 2.34 (nearest even)
- With sticky bit

0.00501	---->	0.00501
2.34000	---->	2.3400
2.34501	---->	2.35 (rounded up from .501)