Chapter - 19
Floating Point
## Floating Point Format

- The sign (plus or minus).
- The 4 digit fraction.
- The single-digit exponent.

Zero is 0.0

We represent these numbers in “E” format: ±f.ffffE±e.

### Examples:

<table>
<thead>
<tr>
<th>Notation</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>+1.000E+ 0</td>
<td>1.0</td>
</tr>
<tr>
<td>+3.300E+ 5</td>
<td>330000</td>
</tr>
<tr>
<td>-8.223E-3</td>
<td>-0.01</td>
</tr>
<tr>
<td>+0.000E+ 0</td>
<td>0.0</td>
</tr>
</tbody>
</table>
Floating Point Add/Sub

1. Start with the numbers:
   
   +2.000E+0    The number is 2.0
   +3.000E-1    The number is 0.3

2. Add guard digits to both numbers
   
   +2.0000E+0    The number is 2.0
   +3.0000E-1    The number is 0.3

3. Shift the number with the smallest exponent to the right one digit and
   numbers match.
   
   +2.0000E+0    The number is 2.0
   +0.3000E-0    The number is 0.3

4. Add the two fractions. The result has the same exponent as the two
   numbers.
   
   +2.0000E+0    The number is 2.0
   +0.3000E-0    The number is 0.3
   +2.3000E+0    Result 2.3
Floating Point Add/Sub

5. Normalize the number by shifting it left or right until there is just one non-zero digit to the left of the decimal point. Adjust the exponent accordingly. A number like +0.1234E+0 would be normalized to +1.2340E-1. Because the number +2.3000E+0 is already normalized we do nothing.

6. Finally, if the guard digit is greater than or equal to 5, round the next digit up; otherwise truncate the number.
   
   +2.3000E+0  Round last digit
   +2.300E+0   Result 2.3

7. For floating-point subtraction, change the sign of the second operand and add.
Multiplication

1. Add the guard digit:
   
   +1.2000E-1  The number is 0.12  
   +1.1000E+1  The number is 11.0

2. Multiply the two fractions and add the exponents.

   (1.2 *  
   +1.2000E-1  The number is 0.12  
   +1.1000E+1  The number is 11.0  
   +1.3200E+0  The result is 1.32

3. Normalize the result. If the guard digit is less than or equal to 5, round the next digit up. Otherwise, truncate the number.

   +1.3200E+0  The number is 1.32
Division

1. Add the guard digit:
   +1.0000E+2   The number is 100.0
   +3.0000E+1   The number is 30.0

2. Divide the fractions, subtract the exponents:
   +1.0000E+2   The number is 100.0
   +3.0000E+1   The number is 30.0
   +0.3333E+1   The result is 3.333

3. Normalize the result:
   +3.3330E+0   The result is 3.333

4. If the guard digit is less than or equal to 5, round the next
   +3.333E+0   The result is 3.333
Overflow and Underflow

\[ 9.000 \times 10^9 \times 9.000 \times 10^9 \]

is:
\[ 8.1 \times 10^{19} \]
That too big for our representation (overflow).

\[ 1.000 \times 10^{-9} \times 1.000 \times 10^{-9} \]

is
\[ 1.0 \times 10^{-18} \]
That’s to small (underflow).
Roundoff Error

1/3 + 1/3 != 2/3

2/3 as floating-point is 6.667E-1
1/3 as floating-point is 3.3333-1
  +3.333E-1
  +3.333E-1
  +6.666E-1 or 0.6666
which is not:
  +6.667E-1
Accuracy

1 - 1/3 - 1/3 - 1/3

1.000E+0
- 3.333E-1
- 3.333E-1
- 3.333E-1
or:

1.000E+0
- 3.333E-1
- 3.333E-1
- 3.333E-1
0.0010E+0 or 1.000E-3

Minimizing error:
• Use double instead of float
• Other techniques are beyond the scope of this course.
Determining Accuracy

int main() {

    ++counter;

}

break;
++counter;

}"
Some older compilers do everything in double.

```c++
float answer, number1, number2;

answer = number1 + number2;
```

C++ must perform the following steps:
1) Convert number1 from single to double precision.
2) Convert number2 from single to double precision.
3) Double precision add.
4) Convert result into single precision and store in answer.

If the variables were of type `double`, C++ would only have to perform the steps:
1) Double precision add.
2) Store result in answer.
## Power Series

\[
\sin (x) = 1 + x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots
\]

\[
\sin(\pi/2)
\]

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$1 \times$</td>
<td>1.571E + 0</td>
<td></td>
</tr>
<tr>
<td>$2 \frac{x^3}{3!}$</td>
<td>6.462E - 1</td>
<td>9.248E - 1</td>
</tr>
<tr>
<td>$3 \frac{x^5}{5!}$</td>
<td>7.974E - 2</td>
<td>1.005E + 0</td>
</tr>
<tr>
<td>$4 \frac{x^7}{7!}$</td>
<td>4.686E - 3</td>
<td>9.998E - 1</td>
</tr>
<tr>
<td>$5 \frac{x^9}{9!}$</td>
<td>1.606E - 4</td>
<td>1.000E + 0</td>
</tr>
<tr>
<td>$6 \frac{x^{11}}{11!}$</td>
<td>3.604E - 6</td>
<td>1.000E + 0</td>
</tr>
</tbody>
</table>
### $\sin(\pi)$

<table>
<thead>
<tr>
<th>Term</th>
<th>Value</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$x^1$</td>
<td>$3.142E+0$</td>
<td></td>
</tr>
<tr>
<td>$x^3/3!$</td>
<td>$5.170E+0$</td>
<td>$-2.028E+0$</td>
</tr>
<tr>
<td>$x^5/5!$</td>
<td>$2.552E-0$</td>
<td>$5.241E-1$</td>
</tr>
<tr>
<td>$x^7/7!$</td>
<td>$5.998E-1$</td>
<td>$-7.570E-2$</td>
</tr>
<tr>
<td>$x^9/9!$</td>
<td>$8.224E-2$</td>
<td>$6.542E-3$</td>
</tr>
<tr>
<td>$x^{11}/11!$</td>
<td>$7.381E-3$</td>
<td>$-8.388E-4$</td>
</tr>
<tr>
<td>$x^{13}/13!$</td>
<td>$4.671E-4$</td>
<td>$-3.717E-4$</td>
</tr>
<tr>
<td>$x^{15}/15!$</td>
<td>$2.196E-5$</td>
<td>$-3.937E-4$</td>
</tr>
<tr>
<td>$x^{17}/17!$</td>
<td>$7.970E-7$</td>
<td>$-3.929E-4$</td>
</tr>
<tr>
<td>$x^{19}/19!$</td>
<td>$2.300E-8$</td>
<td>$-3.929E-4$</td>
</tr>
</tbody>
</table>