Floating Point

Just superficial coverage

- Scientific notation: decimal vs. binary
  - "Normalized" - one digit to the left of the point
- Floating pt representation:

<table>
<thead>
<tr>
<th>Sign-bit</th>
<th>Exponent</th>
<th>Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>18</td>
<td>23</td>
</tr>
</tbody>
</table>

IEEE 32-bit floating pt number

64-bit

- Leading 1 of significant is implicit
  - All numbers except 0 have it
- In order to represent negative exponents, a bias is used
  \( = 127 \) in 32-bit numbers
- Actual exponent = Exponent field - 127
  - So, for \( 2^{-2} \), exponent field would be 129
- 0 is represented by all 0's
  - There is no \( 0 \times 2^{-127} \) representable.
Note: Not using $2^5$-complement in the exponent or for the significant allows comparison between floating pt numbers using the entire word (since sign is most significant, then exponent, then significant).

FP Addition (quick review)

- Need the operands to have the same exponent.
  - If not, "de-normalize" the smaller number by shifting the significant to the right

- Add the significands (remember the implicit leading ones)

- Re-normalize *watch out for underflow or overflow*

FP Multiplication:

- Add the exponents
  - Subtract or bias (otherwise summed)

- Multiply significands
- Re-normalize product

*Watch out for underflow or overflow*