Topics of this Slideset

IEEE Floating Point Standard
- Rounding
- Floating point operations
- Mathematical properties

IEEE Standard 754
- Established in 1985 as a uniform standard for floating point arithmetic
- It is supported by all major CPUs.
- Before 1985 there were many idiosyncratic formats.

Driven by Numerical Concerns
- Nice standards for rounding, overflow, underflow
- Hard to make go fast: numerical analysts predominated over hardware types in defining the standard
- Now all (add, subtract, multiply) operations are fast except divide.

Floating Point Puzzles

For each of the following, either:
- argue that it is true for all argument values, or
- explain why it is not true.
Assume neither d nor f is NaN.

```
int x = ...;
float f = ...;
double d = ...;
```

- $x == (int)(float) x$
- $x == (int)(double) x$
- $f == (float)(double) f$
- $d == (float) d$
- $f == (-f)$
- $2/3 == 2/3.0$
- $d < 0.0 \rightarrow ((d+2) < 0.0)$
- $d > f \rightarrow -f > -d$
- $d*d >= 0.0$
- $(d+f)-d == f$

IEEE Floating Point Standard

Established in 1985 as a uniform standard for floating point arithmetic
- It is supported by all major CPUs.
- Before 1985 there were many idiosyncratic formats.

Driven by Numerical Concerns
- Nice standards for rounding, overflow, underflow
- Hard to make go fast: numerical analysts predominated over hardware types in defining the standard
- Now all (add, subtract, multiply) operations are fast except divide.
The binary number \( b_i b_{i-1} b_2 b_1 \ldots b_0 . b_{-1} b_{-2} b_{-3} \ldots b_{-j} \) represents a particular (positive) sum. Each digit is multiplied by a power of two according to the following chart:

<table>
<thead>
<tr>
<th>Bit:</th>
<th>( b_i )</th>
<th>( b_{i-1} )</th>
<th>\ldots</th>
<th>( b_2 )</th>
<th>( b_1 )</th>
<th>( b_0 )</th>
<th>( b_{-1} )</th>
<th>( b_{-2} )</th>
<th>( b_{-3} )</th>
<th>\ldots</th>
<th>( b_{-j} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weight:</td>
<td>( 2^i )</td>
<td>( 2^{i-1} )</td>
<td>\ldots</td>
<td>( 2^1 )</td>
<td>( 1 )</td>
<td>( 1/2 )</td>
<td>( 1/4 )</td>
<td>( 1/8 )</td>
<td>( 1/16 )</td>
<td>\ldots</td>
<td>( 2^{-j} )</td>
</tr>
</tbody>
</table>

**Representation:**
- Bits to the right of the *binary point* represent fractional powers of 2.
- This represents the rational number:

\[
\sum_{k=-j}^{i} b_k \times 2^k
\]

The sign is treated separately.

### Representable Numbers

<table>
<thead>
<tr>
<th>Value</th>
<th>Representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 + 3/4</td>
<td>101.11_2</td>
</tr>
<tr>
<td>2 + 7/8</td>
<td>10.111_2</td>
</tr>
<tr>
<td>63/64</td>
<td>0.1111111_2</td>
</tr>
</tbody>
</table>

**Observations**
- Divide by 2 by shifting right
- Multiply by 2 by shifting left
- Numbers of the form 0.1111\ldots_2 are just below 1.0
  - \( 1/2 + 1/4 + 1/8 + \ldots + 1/2^i \rightarrow 1.0 \)
  - We use the notation \( 1.0 - \epsilon \).

### Limitation
- You can only represent numbers of the form \( y + x/2^i \).
- Other fractions (rationals) have repeating bit representations
- Irrationals have infinite, non-repeating representations
Aside: Converting Decimal Fractions to Binary

If you want to convert a decimal fraction to binary, it's easy if you use a simple iterative procedure.

1. Start with the decimal fraction (> 1) and multiply by 2.
2. Stop if the result is 0 (terminated binary) or a result you’ve seen before (repeating binary).
3. Record the whole number part of the result.
4. Repeat from step 1 with the fractional part of the result.

Let's try another one, 0.1 or 1/10

0.1 * 2 = 0.2
0.2 * 2 = 0.4
0.4 * 2 = 0.8
0.8 * 2 = 1.6
0.6 * 2 = 1.2
0.2 * 2 = 0.4

We could continue, but we see that it's going to repeat forever (since 0.2 repeats our multiplicand from the second line). Reading the integer parts from the top gives 0[0011], since we'll repeat the last 4 bits forever.

Aside: Converting Decimal Fractions to Binary (2)

The result (following the binary point) is the series of whole numbers components of the answers read from the top, i.e., 0.011.

Floating Point Representation

Numerical Form

\[-1^s \times M \times 2^E\]

- Sign bit \( s \) determines whether number is negative or positive.
- Significand \( M \) is normally a fractional value in the range \([1.0 \ldots 2.0]\)
- Exponent \( E \) weights value by power of two.

Encoding

<table>
<thead>
<tr>
<th>s</th>
<th>exp</th>
<th>frac</th>
</tr>
</thead>
</table>

- The most significant bit is the sign bit.
- The exp field encodes \( E \).
- The frac field encodes \( M \).

Float format:
Floating Point Precisions

Encoding

| s | exp | frac |

- The most significant bit is the sign bit.
- The exp field encodes E.
- The frac field encodes M.

Sizes

- Single precision: 8 exp bits, 23 frac bits, for 32 bits total
- Double precision: 11 exp bits, 52 frac bits, for 64 bits total
- Extended precision: 15 exp bits, 63 frac bits (only Intel-compatible machines)

Normalized Numeric Values

Condition: \( \exp \neq 000\ldots0 \) and \( \exp \neq 111\ldots1 \)

Exponent is coded as a biased value

\[ E = \text{Exp} - \text{Bias} \]

- \text{Exp}: unsigned value denoted by exp.
- \text{Bias}: Bias value
  - In general: \( \text{Bias} = 2^{\text{Exp} - 1} - 1 \), where \( e \) is the number of exponent bits
  - Single precision: 127 (Exp: 1...254, E: −126...127)
  - Double precision: 1023 (Exp: 1...2046, E: −1022...1023)

Significand coded with implied leading 1

\[ M = 1.xxxx \ldots x_2 \]

- \( xxx\ldots x \): bits of frac
- Minimum when 000...0 (\( M = 1.0 \))
- Maximum when 111...1 (\( M = 2.0 - \epsilon \))
- We get the extra leading bit “for free.”

Converting Between Float and Decimal

Value:

\[
\text{float } F = 15213.0; \\

15213_{10} = 11101101101101_{2} = 1.1101101101101_{2} \times 2^{13}
\]

Significand

\( M = 1.1101101101101_{2} \)
\( \text{frac} = 11011011011010000000000000 \)

Exponent

\( E = 13 \)
\( \text{Bias} = 127 \)
\( \text{Exp} = 140 = 10001100 \)
Given the bit string 0x40500000, what floating point number does it represent?

Writing this as a bit string gives us:

0 10000000 10100000000000000000000

We see that this is a positive, normalized number.

exp = 128 − 127 = 1

So, this number is:

1.1011 × 2^1 = 11.011_2 = 3.25_{10}

Denormalized Values

Condition: \( \text{exp} = 000\ldots0 \)

Value

- Exponent values: \( E = -\text{Bias} + 1 \) Why this value?
  - Floats: −126; Doubles: −1022
- Significand value: \( M = 0.xxx\ldots x_2 \), where \( xxx\ldots x \) are the bits of frac.

Cases

- \( \text{exp} = 000\ldots0 \) and \( \text{frac} = 000\ldots0 \)
  - represents values of 0
  - notice that we have distinct +0 and -0
- \( \text{exp} = 000\ldots0 \) and \( \text{frac} \neq 000\ldots0 \)
  - These are numbers very close to 0.0
  - Lose precision as they get smaller
  - Experience “gradual underflow”
Given the bit string 0x80600000, what floating point number does it represent?

Writing this as a bit string gives us:

```
1 00000000 11000000000000000000000
```

We see that this is a negative, denormalized number with value:

\[-0.11_2 \times 2^{-126} = -1.1_2 \times 2^{-127}\]

The smallest positive normalized float is \(1.0_2 \times 2^{-126}\). Why did I get that number? All positive normalized floats are greater or equal.

The largest positive denormalized float is \(0.11111111111111111_2 \times 2^{-126}\). Why? All positive denorms are between this number and 0.

Note that the smallest norm and the largest denorm are incredibly close together. How close? Thus, the normalized range flows naturally into the denormalized range because of this choice of exponent for denorms.

**Condition:** \(\exp = 111...1\)

**Cases**

- \(\exp = 111...1\) and \(\text{frac} = 000...0\)
  - Represents value of infinity (\(\infty\))
  - Result returned for operations that overflow
  - Sign indicates positive or negative
  - E.g., \(1.0/0.0 = -1.0/ - 0.0 = +\infty, \ 1.0/ - 0.0 = -\infty\)

- \(\exp = 111...1\) and \(\text{frac} \neq 000...0\)
  - Not-a-Number (NaN)
  - Represents the case when no numeric value can be determined
  - E.g., \(\sqrt{-1}, \infty - \infty\)

How many 32-bit NaN’s are there?
8-bit Floating Point Representation

- The sign bit is in the most significant bit.
- The next four bits are the exponent with a bias of 7.
- The last three bits are the frac.

This has the general form of the IEEE Format

- Has both normalized and denormalized values.
- Has representations of 0, NaN, infinity.

\[
\begin{array}{cccc}
7 & 6 & 5 & 4 \\
& \text{exp} & \text{frac} & \\
\end{array}
\]

Values Related to the Exponent

<table>
<thead>
<tr>
<th>Exp</th>
<th>exp</th>
<th>E</th>
<th>$2^E$</th>
<th>comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>000</td>
<td>-6</td>
<td>1/64</td>
<td>(denorms)</td>
</tr>
<tr>
<td>1</td>
<td>000</td>
<td>-6</td>
<td>1/64</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>000</td>
<td>-5</td>
<td>1/32</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>000</td>
<td>-4</td>
<td>1/16</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>001</td>
<td>-3</td>
<td>1/8</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>010</td>
<td>-2</td>
<td>1/4</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>011</td>
<td>-1</td>
<td>1/2</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>111</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>100</td>
<td>+1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>100</td>
<td>+2</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>101</td>
<td>+3</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>101</td>
<td>+4</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>110</td>
<td>+5</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>110</td>
<td>+6</td>
<td>64</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>110</td>
<td>+7</td>
<td>128</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>111</td>
<td>n/a</td>
<td></td>
<td>(inf, NaN)</td>
</tr>
</tbody>
</table>

Dynamic Range

Simple Float System

Notice that the floating point numbers are not distributed evenly on the number line.

Suppose $M$ is the largest possible exponent, $m$ is the smallest, $\frac{1}{8}$ is the smallest positive number representable, and $\frac{7}{4}$ the largest positive number representable. What is the format?
**Interesting FP Numbers**

<table>
<thead>
<tr>
<th>Description</th>
<th>exp</th>
<th>frac</th>
<th>Numeric value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Zero</td>
<td>00...00</td>
<td>00...00</td>
<td>0.0 (2^{(-23,-52)} \times 2^{(-126,-1022)})</td>
</tr>
<tr>
<td>Smallest Pos. Denorm</td>
<td>00...00</td>
<td>00...01</td>
<td>(1.0 - \epsilon) (2^{(-126,-1022)})</td>
</tr>
<tr>
<td>Largest Denorm</td>
<td>00...00</td>
<td>11...11</td>
<td>(1.0 \times 2^{(-126,-1022)})</td>
</tr>
<tr>
<td>Smallest Pos. Norm.</td>
<td>00...01</td>
<td>00...00</td>
<td>1.0 (2^{(127,1023)})</td>
</tr>
<tr>
<td>One</td>
<td>01...11</td>
<td>00...00</td>
<td>1.0</td>
</tr>
<tr>
<td>Largest Norm.</td>
<td>11...10</td>
<td>11...11</td>
<td>(2.0 - \epsilon) (2^{(127,1023)})</td>
</tr>
</tbody>
</table>

**Float Point Operations**

**Conceptual View**
- First compute the exact result.
- Make it fit into the desired precision.
  - Possibly overflows if exponent is too large.
  - Possibly round to fit into frac.

**Rounding Modes (illustrated with $ rounding)**

<table>
<thead>
<tr>
<th></th>
<th>$1.40</th>
<th>$1.60</th>
<th>$1.50</th>
<th>$2.50</th>
<th>-$1.50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Toward Zero</td>
<td>$1</td>
<td>$1</td>
<td>$1</td>
<td>$2</td>
<td>-$1</td>
</tr>
<tr>
<td>Round down (-\infty)</td>
<td>$1</td>
<td>$1</td>
<td>$1</td>
<td>$2</td>
<td>-$2</td>
</tr>
<tr>
<td>Round up (+\infty)</td>
<td>$2</td>
<td>$2</td>
<td>$2</td>
<td>$3</td>
<td>-$1</td>
</tr>
<tr>
<td>Nearest even (default)</td>
<td>$1</td>
<td>$2</td>
<td>$2</td>
<td>$2</td>
<td>-$2</td>
</tr>
</tbody>
</table>

- Round down: rounded result is close to but no greater than true result.
- Round up: rounded result is close to but no less than true result.

**Closer Look at Round to Even**

**Default Rounding Mode**
- Hard to get any other kind without dropping into assembly.
- All others are statistically biased; the sum of a set of integers will consistently be under- or over-estimated.

**Applying to Other Decimal Places / Bit Positions**
When exactly halfway between two possible values, round so that the least significant digit is even.

E.g., round to the nearest hundredth:

- 1.2349999 \(\approx 1.23\) Less than half way
- 1.2350000 \(\approx 1.24\) Greater than half way
- 1.2350001 \(\approx 1.24\) Half way, round up
- 1.2450000 \(\approx 1.24\) Half way, round down

**FP Zero is the Same as Integer Zero:** All bits are 0.

**Can (Almost) Use Unsigned Integer Comparison**
- Must first compare sign bits.
- Must consider \(-0 = 0\).
- NaNs are problematic:
  - Will be greater than any other values.
  - What should the comparison yield?
- Otherwise, it’s OK.
  - Denorm vs. normalized works.
  - Normalized vs. infinity works.
Binary Fractional Numbers
- “Even” when least significant bit is 0.
- Half way when bits to the right of rounding position = 10[0]₂.

Examples
E.g., Round to nearest 1/4 (2 bits to right of binary point).

<table>
<thead>
<tr>
<th>Value</th>
<th>Binary</th>
<th>Rounded</th>
<th>Action</th>
<th>Rounded Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 3/32</td>
<td>10.00011₂</td>
<td>10.00</td>
<td>(&lt; 1/2: down)</td>
<td>2</td>
</tr>
<tr>
<td>2 3/16</td>
<td>10.00110₂</td>
<td>10.01</td>
<td>(&gt; 1/2: up)</td>
<td>2 1/4</td>
</tr>
<tr>
<td>2 7/8</td>
<td>10.11100₂</td>
<td>11.00</td>
<td>(1/2: up)</td>
<td>3</td>
</tr>
<tr>
<td>2 5/8</td>
<td>10.10100₂</td>
<td>10.10</td>
<td>(1/2: down)</td>
<td>2 1/2</td>
</tr>
</tbody>
</table>

Round to Even
When rounding to even, consider the two possible choices and choose the one with a 0 in the final position.

Example: round to nearest 1/4 using round to even:

- 1.10 → 1 1/2
- 1.1010000 → 1 5/8
- 1.11 → 1 3/4
- 1.1110000 → 1 7/8
- 10.00 → 2.0

FP Multiplication

Operands: \((-1)^{S_1} \times M_1 \times 2^{E_1}, (-1)^{S_2} \times M_2 \times 2^{E_2}\)

Exact Result: \((-1)^S \times M \times 2^E\)
- Sign S: \(S_1\) xor \(S_2\)
- Significand M: \(M_1 \times M_2\)
- Exponent E: \(E_1 + E_2\)

Fixing
- If \(M \geq 2\), shift M right, increment E
- E is out of range, overflow
- Round M to fit frac precision

Implementation
Biggest chore is multiplying significands.
**Binary Example**

\[(-1.01 \times 2^2)(1.1 \times 2^4)\]

\[= (-1.01 \times 1.1)(2^2 \times 2^4)\]

\[= -1.111 \times 2^6\]

\[= -10.0 \times 2^6\]

\[= -1.0 \times 2^7\]

round to even

adjust exponent

Be careful if you try to do this in the floating point format, rather than in scientific notation. Since the exponents are biased in FP format, adding them would give you:

\[(2 + \text{bias}) + (4 + \text{bias}) = 6 + 2*\text{bias}\]

To adjust you have to subtract the bias.

**FP Addition**

**Operands:** \((-1)^{S_1} \times M_1 \times 2^{E_1}, (-1)^{S_2} \times M_2 \times 2^{E_2}\)

Assume \(E_1 > E_2\)

**Exact Result:** \((-1)^S \times M \times 2^E\)

- Sign \(S\), Significand \(M\); result of signed align and add.
- Exponent \(E\): \(E_1\)

**Fixing**

- If \(M \geq 2\), shift \(M\) right, increment \(E\)
- If \(M < 1\), shift \(M\) left \(k\) positions, decrement \(E\) by \(k\)
- if \(E\) is out of range, overflow
- Round \(M\) to fit frac precision

*If you try to do this in the FP form, recall that both exponents are biased.*

**Mathematical Properties of FP Add**

**Compare to those of Abelian Group**

- Closed under addition? Yes, but may generate infinity or NaN.
- Commutative? Yes.
- Associative? No, because of overflow and inexactness of rounding.
- \(O\) is additive identity? Yes.
- Every element has additive inverse? Almost, except for infinities and NaNs.

**Monotonicity**

- \(a \geq b \implies a + c \geq b + c\)? Almost, except for infinities and NaNs.
Mathematical Properties of FP Mult

**Compare to those of Commutative Ring**
- Closed under multiplication? Yes, but may generate infinity or NaN.
- Multiplication Commutative? Yes.
- Multiplication is Associative? No, because of possible overflow and inexactness of rounding.
- 1 is multiplicative identity? Yes.
- Multiplication distributes over addition? No, because of possible overflow and inexactness of rounding.

**Monotonicity**
\[ a \geq b \& c \geq 0 \implies a \times c \geq b \times c? \text{ Almost, except for infinities and NaNs.} \]

Floating Point in C

**C guarantees two levels**
- float: single precision
- double: double precision

**Conversions**
- Casting among int, float, and double changes numeric values
- Double or float to int:
  - truncates fractional part
  - like rounding toward zero
  - not defined when out of range: generally saturates to TMin or TMax
- int to double: exact conversion as long as int has \leq 53-bit word size
- int to float: will round according to rounding mode.

Answers to FP Puzzles

```c
int x = ...;
float f = ...;
double d = ...;
```

Assume neither d nor f is NaN.

\[
\begin{align*}
x & \approx (\text{int})(\text{float}) x \\
x & \approx (\text{int})(\text{double}) x \\
f & \approx (\text{float})(\text{double}) f \\
d & \approx (\text{float}) d \\
f & \approx (-f) \\
2/3 & \approx 2/3.0 \\
d & < 0.0 \implies ((d \times 2) < 0.0) \\
d & > f \implies -f > -d \\
d \times d & \approx 0.0 \\
(d + f) - d & \approx f
\end{align*}
\]

- No: 24 bit significand
- Yes: 53 bit significand
- Yes: increases precision
- No: loses precision
- Yes: just change sign bit
- No: 2/3 \approx 0
- Yes
- Yes
- Yes
- No: not associative

Ariane 5

On June 4, 1996 an unmanned Ariane 5 rocket launched by the European Space Agency exploded just forty seconds after its lift-off from Kourou, French Guiana. The rocket was on its first voyage, after a decade of development costing \$7 billion.

The destroyed rocket and its cargo were valued at \$500 million.
The cause of the failure was a software error in the inertial reference system.

Specifically a 64-bit floating point number relating to the horizontal velocity of the rocket with respect to the platform was converted to a 16-bit signed integer.

The number was larger than 32,767, the largest integer storeable in a 16-bit signed integer, and thus the conversion failed.

IEEE Floating Point has Clear Mathematical Properties

- Represents numbers of the form $\pm M \times 2^E$.
- Can reason about operations independent of implementation: as if computed with perfect precision and then rounded.
- Not the same as real arithmetic.
  - Violates associativity and distributivity.
  - Makes life difficult for compilers and serious numerical application programmers.