Ints and Floats

15-213 Spring 2007

Aleksey Kliger

Computer Science Department
Carnegie Mellon University

January 29, 2007

Announcements

- PLEASE log into the fish machines TODAY.
  - Especially if you've never used unix before
- First quiz goes Blackboard today
  - 30 minutes without breaks
  - Open book, notes
  - No Google, Wikipedia, or writing programs
  - If you lose net connection or your computer explodes, email `fp@cs`

Outline

Shift left

- $x \ll n$ defined iff $0 \leq n < bits(typeof(x))$.
- If x is an `int`, $x \ll 32$ is undefined.
  - No error.
  - Different answers on different machines.
- For Datalab: If it passes the BDD checker, it's correct.

Integers

Unsigned

$$0, \ldots, 2^n - 1$$

Signed

$$-2^{n-1}, \ldots, 2^{n-1} - 1$$

Properties

- $+, *$ are associative, commutative
- Not always the case: $x + y \geq x$

Outline

`unsigned char`

Example

- $0xA3 + 0xA7 = ??$

$$\begin{array}{rcc} & 1010 & 0011 \\ + & 1010 & 0111 \\ \hline \end{array}$$

`unsigned char`

Example

- $0xA3 + 0xA7 = 0x4A$

$$\begin{array}{rrr} & 1010 & 0011 \\ + & 1010 & 0111 \\ \hline & 0100 & 1010 \end{array}$$

$$163 + 167 = 74$$

`unsigned char`

Example

- $0xA3 + 0xA7 = 0x4A$

$$\begin{array}{rll} & 1010 & 0011 \\ + & 1010 & 0111 \\ \hline & 0100 & 1010 \end{array}$$

$$163 + 167 = 74$$

- $0xFF + 0x01 = 0x00$

Outline

1. Announcements
2. A Note On Shifting
3. Integer Representations
   - Unsigned Ints
   - Signed Ints
4. IEEE Floating Point
   - Representation
   - Rounding
   - Conversions and overflow in C

Signed integers

- MSB is 1 iff negative
- To negate: flip the bits and add one.
- `0x80000000` is most negative `int`
- `0xFFFFFFFF` is -1
- `0x7FFFFFFF` is most positive `int`

`signed char`

Example

- $0xFF + 0x01 = 0x00$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
  - Still $0x4A$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
  - Still $0x4A$
  - But different interpretation:

$$\tilde{}0xA3 + 1 =$$

$$\tilde{}0xA7 + 1 =$$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
  - Still $0x4A$
  - But different interpretation:

$$\tilde{}0xA3 + 1 = \tilde{}(1010\ 0011) + 1 = 0101\ 1100 + 1 =$$

$$\tilde{}0xA7 + 1 =$$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 =??$
  - Still $0x4A$
  - But different interpretation:

    $$\tilde{}0xA3 + 1 = \tilde{}(1010\ 0011) + 1 = 0101\ 1100 + 1 = 0101\ 1101 = 0x5D$$

    So $0xA3$ means -93

    $$\tilde{}0xA7 + 1 =$$

signed char

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
    - Still $0x4A$
    - But different interpretation:

$$\tilde{}0xA3 + 1 = \tilde{}(1010\ 0011) + 1 = 0101\ 1100 + 1 = 0101\ 1101 = 0x5D$$

So $0xA3$ means -93

$$\tilde{}0xA7 + 1 = \tilde{}(1010\ 0111) + 1 = 0101\ 1000 + 1 = 0101\ 1001 = 0x59$$

`signed char`

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
  - Still $0x4A$
  - But different interpretation:

    $$\tilde{}0xA3 + 1 = \tilde{}(1010\ 0011) + 1 = 0101\ 1100 + 1 = 0101\ 1101 = 0x5D$$

    So $0xA3$ means -93

    $$\tilde{}0xA7 + 1 = \tilde{}(1010\ 0111) + 1 = 0101\ 1000 + 1 = 0101\ 1001 = 0x59$$

    So $0xA7$ means -89

`signed char`

Example

- $0xFF + 0x01 = 0x00$
- $0xA3 + 0xA7 = ??$
  - Still $0x4A$
  - But different interpretation:

    $$\tilde{}0xA3 + 1 = \tilde{}(1010\ 0011) + 1 = 0101\ 1100 + 1 = 0101\ 1101 = 0x5D$$

    So $0xA3$ means -93

    $$\tilde{}0xA7 + 1 = \tilde{}(1010\ 0111) + 1 = 0101\ 1000 + 1 = 0101\ 1001 = 0x59$$

    So $0xA7$ means -89
  - $-93 + -89 = -182 = 74$

Multiplication

In general multiplying two n-bit numbers can give a $2n$-bit result

$$0x80 \times 0x80 = 0x4000$$

$$128 \times 128 = 2^{14} = 16384$$

Outline

Floating point numbers

- Approximate a finite subset of $\mathbb{R}$
- Doesn't behave like $\mathbb{R}$
  - Not associative
  - Not always: $x = y$ or $x < y$ or $y < x$
- Multiple representations

normalized	most numbers
denormalized	very small numbers
other	Infinities, NaN

Base 2 fractions

- Analog to base 10:

$$0.23 = 2 \times 10^{-1} + 3 \times 10^{-2}$$

- 𝅗𝅥, ♩, ♪
- Can represent some rationals exactly:

$$\frac{3}{8} = 0.011_2$$

- Others need infinite precision:

$$\frac{1}{5} = 0.00110011\ldots$$

Normalized Representation

$(-1)^s \times 1.M \times 2^{E-Bias}$

- 1-bit sign, 1 iff negative
- n-bit exponent in *biased form*
  - excluding $0 \cdots 0, 1 \cdots 1$
  - $Bias = 2^{n-1} - 1$
- k-bit significand as *implied leading 1 fractional part*

Example (3-bit exponent, 4-bit significand)

0 101 0111

Denormalized Representation

$(-1)^s \times 0.M \times 2^{(1-Bias)}$

- 1-bit sign
- n-bit exponent set to zero
- $Bias = 2^{n-1} - 1$
- k-bit significand as fractional part

Example

1 000 0000

0 000 1110

Infinity and NaN

$(-1)^s \times \infty$

- 1-bit sign
- n-bit exponent is all ones
- k-bit significand is all zeroes
- Bigger (smaller) than all other floats
- Two infinities of the same sign are equal

Example

1 111 0000

NaN

- 1-bit sign
- n-bit exponent is all ones
- k-bit significand is not all zeroes
- Incomparable to all floats, including to itself
- $0.0/0.0$, $\infty - \infty$

Example

0 111 0110

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00				
0 00 01				
0 00 10				
0 00 11				
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm			
0 00 01				
0 00 10				
0 00 11				
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$		
0 00 01				
0 00 10				
0 00 11				
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	
0 00 01				
0 00 10				
0 00 11				
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01				
0 00 10				
0 00 11				
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0		
0 00 10	Denorm	0		
0 00 11	Denorm	0		
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	
0 00 10	Denorm	0		
0 00 11	Denorm	0		
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0		
0 00 11	Denorm	0		
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0		
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00				
0 01 01				
0 01 10				
0 01 11				
0 10 00				
...				
0 10 11				
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - \textit{Bias} = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm			
0 01 01	Norm			
0 01 10	Norm			
0 01 11	Norm			
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$		
0 01 01	Norm	0		
0 01 10	Norm	0		
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	
0 01 01	Norm	0		
0 01 10	Norm	0		
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0		
0 01 10	Norm	0		
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	
0 01 10	Norm	0		
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - \mathit{Bias} = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - \mathit{Bias} = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0		
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0		
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm			
...				
0 10 11	Norm			
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$		
...				
0 10 11	Norm	1		
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$	0	$(1 + 0) \times 2^1 = 2$
...				
0 10 11	Norm	1		
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$	0	$(1 + 0) \times 2^1 = 2$
...				
0 10 11	Norm	1	$\frac{3}{4}$	
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$	0	$(1 + 0) \times 2^1 = 2$
...				
0 10 11	Norm	1	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^1 = 3.5$
0 11 00				
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$	0	$(1 + 0) \times 2^1 = 2$
...				
0 10 11	Norm	1	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^1 = 3.5$
0 11 00	Infty	—	—	$+\infty$
0 11 01				

Example (5-bit float): 2-bit exponent, 2-bit significand

$$\text{Bias} = 2^{2-1} - 1 = 1$$

Bits	Rep'n	Exponent	Signif.	Value
0 00 00	Denorm	$1 - Bias = 0$	0	$+0.0$
0 00 01	Denorm	0	$\frac{1}{4}$	$(0 + \frac{1}{4}) \times 2^0 = 0.25$
0 00 10	Denorm	0	$\frac{1}{2}$	$(0 + \frac{1}{2}) \times 2^0 = 0.5$
0 00 11	Denorm	0	$\frac{1}{2} + \frac{1}{4}$	$(0 + \frac{3}{4}) \times 2^0 = 0.75$
0 01 00	Norm	$1 - Bias = 0$	0	$(1 + 0) \times 2^0 = 1.0$
0 01 01	Norm	0	$\frac{1}{4}$	$(1 + \frac{1}{4}) \times 2^0 = 1.25$
0 01 10	Norm	0	$\frac{1}{2}$	$(1 + \frac{1}{2}) \times 2^0 = 1.5$
0 01 11	Norm	0	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^0 = 1.75$
0 10 00	Norm	$2 - Bias = 1$	0	$(1 + 0) \times 2^1 = 2$
...				
0 10 11	Norm	1	$\frac{3}{4}$	$(1 + \frac{3}{4}) \times 2^1 = 3.5$
0 11 00	Infty	—	—	$+\infty$
0 11 01	NaN	—	—	NaN

Outline

Rounding

$x \odot y$

1. Convert to $\mathbb{R}$
2. Perform exact operation
3. Round result to fit in a `float` (`double`)

Rounding modes

Mode	1.5	-1.5	1.4	1.6
Down	1	-2	1	1
Up	2	-1	2	2
To zero	1	-1	1	1
To even	2	-2	1	2

Round to even

- Avoids statistical mistakes of other rounding modes
- On average half round up, half round down
- Round to nearest representable value
- If exactly halfway between, round to LSB $==0$

Outline

Conversions in C

- Are there enough bits?
  - `float`: 23-bit significand, 8-bit exponent
  - `double`: 52-bit significand, 11-bit exponent
  - `int` has 32 bits, all of them important
- Exact
  - int to double, float to double
- May lose precision
  - int to float
- May overflow and lose precision
  - double to float, anything to int