CSSE132
Introduction to Computer Systems

5 : Floating point

March 11, 2013

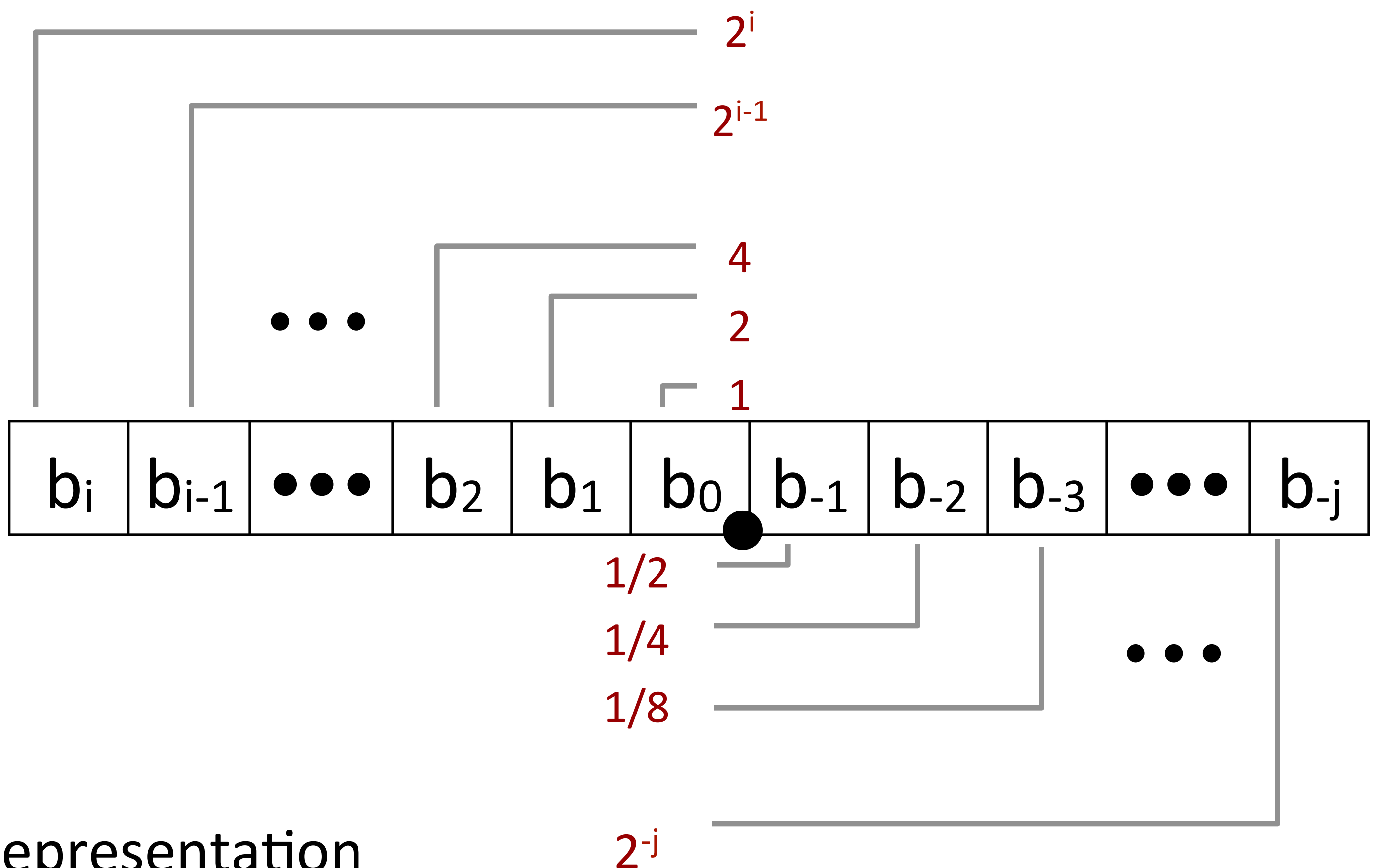
Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point format
- Examples
- Basic conversion
- Properties

Fractional binary numbers

- What is 1011.101_2?

Fractional Binary Numbers



- Representation
  - Bits to right of "binary point" represent fractional powers of 2
  - Represents rational number:

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

Fractional Binary Numbers

0	0	1	0	0	1	0	1	bit
1/2	1/4	1/8	1/16	1/32	1/64	1/128	1/256	Place value
2^{-1}	2^{-2}	2^{-3}	2^{-4}	2^{-5}	2^{-6}	2^{-7}	2^{-8}	2^{-n} value

1/8 + 1/64 + 1/256 = 32/256 + 4/256 + 1/256 = 37/256 = 0.14453125

- Each bit is a negative power of 2
  - $2^{-1} = 1/2$
  - $2^{-2} = 1/2^2$
  - …

Fractional Binary Numbers: Examples

- Value / Representation

Value	Representation
5 3/4	101.11_2
2 7/8	10.111_2
2 1/2	10.1_2
3 1/4	11.01_2

- Observations
  - Divide by 2 by shifting right
  - Multiply by 2 by shifting left
  - Numbers of form $0.111111\ldots_2$ are just below 1.0
    - $1/2 + 1/4 + 1/8 + \ldots + 1/2^i + \ldots \rightarrow 1.0$
    - Use notation $1.0 - \varepsilon$

Representable Numbers

- Limitation
  - Can only exactly represent numbers of the form $x/2^k$
  - Other rational numbers have repeating bit representations

- Value / Representation
  - 1/3 — $0.0101010101[01]\ldots_2$
  - 1/5 — $0.001100110011[0011]\ldots_2$
  - 1/10 — $0.0001100110011[0011]\ldots_2$

Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point format
- Examples
- Basic conversion
- Properties

IEEE Floating Point

- IEEE Standard 754
  - Established in 1985 as uniform standard for floating point arithmetic
    - Before that, many idiosyncratic formats
  - Supported by all major CPUs

- Driven by numerical concerns
  - Nice standards for rounding, overflow, underflow
  - Hard to make fast in hardware
    - Numerical analysts predominated over hardware designers in defining standard

Floating Point Representation

- Numerical Form:

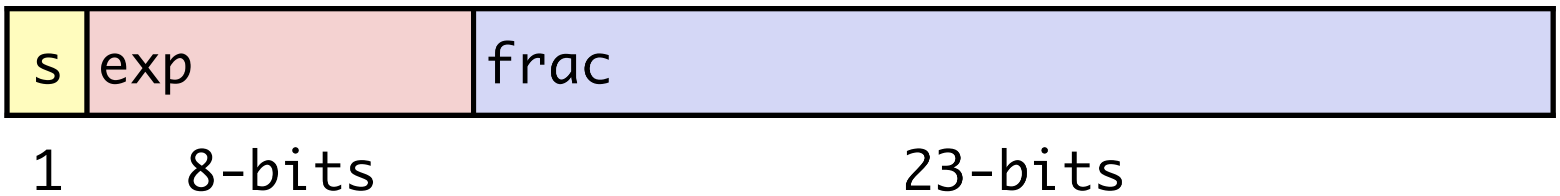
$$(-1)^s \, M \;\; 2^E$$

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

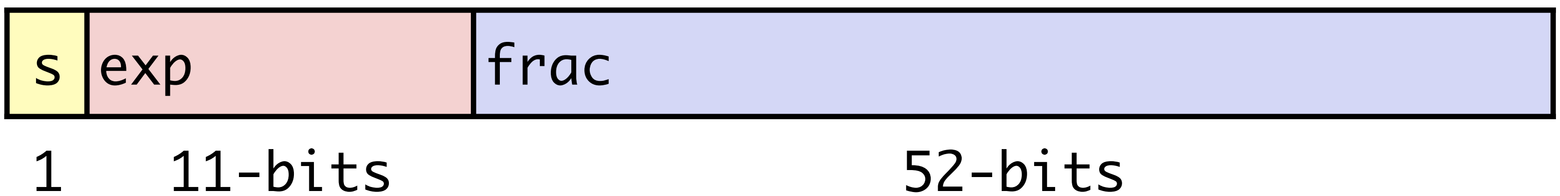
- Encoding
  - MSB `s` is sign bit s
  - `exp` field encodes E (but is not equal to E)
  - `frac` field encodes M (but is not equal to M)

Precisions

- Single precision: 32 bits



- Double precision: 64 bits



- Also quad precision (128 bit) and half precision (16 bit)

Normalized Values

- Condition: exp ≠ 000…0 and exp ≠ 111…1

- Exponent coded as biased value: E = Exp – Bias
  - Exp: unsigned value `exp`
  - Bias = $2^{k-1} - 1$, where k is 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 = $1.xxx…x_2$
  - xxx…x: bits of `frac`
  - Minimum when `000…0` (M = 1.0)
  - Maximum when `111…1` (M = 2.0 – ε)
  - Get extra leading bit for "free"

Normalized Encoding Example

- **Value:** `Float F = 15213.0;`
  - $15213_{10} = 11101101101101_2$
    $= 1.1101101101101_2 \times 2^{13}$

- **Significand**

  M = `1.`$\underline{1101101101101}_2$

  `frac=` $\underline{1101101101101}0000000000_2$

- **Exponent**

  E = 13

  $Bias$ = 127

  Exp = 140 = 10001100_2

- **Result:**

s	exp	frac
0	10001100	11011011011010000000000

Denormalized Values

- Condition: `exp = 000…0`

- Exponent value: E = –Bias + 1 (instead of E = 0 – Bias)
- Significand coded with implied leading 0: $M = 0.xxx\ldots x_2$
  - `xxx…x`: bits of `frac`
- Cases
  - `exp` = 000…0, `frac` = 000…0
    - Represents zero value
    - Note distinct values: +0 and –0 (why?)
  - `exp` = 000…0, `frac` ≠ 000…0
    - Numbers very close to 0.0
    - Lose precision as get smaller
    - Equispaced

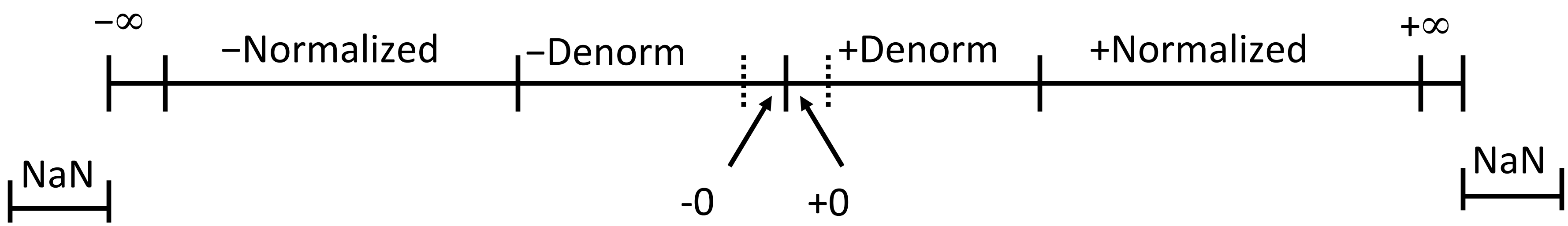
Special Values

- Condition: `exp` = 111…1

- Case: `exp` = 111…1, `frac` = 000…0
  - Represents value ∞ (infinity)
  - Operation that overflows
  - Both positive and negative
  - E.g., 1.0/0.0 = −1.0/−0.0 = +∞, 1.0/−0.0 = −∞

- Case: `exp` = 111…1, `frac` ≠ 000…0
  - Not-a-Number (NaN)
  - Represents case when no numeric value can be determined
  - E.g., sqrt(–1), ∞ − ∞, ∞ × 0

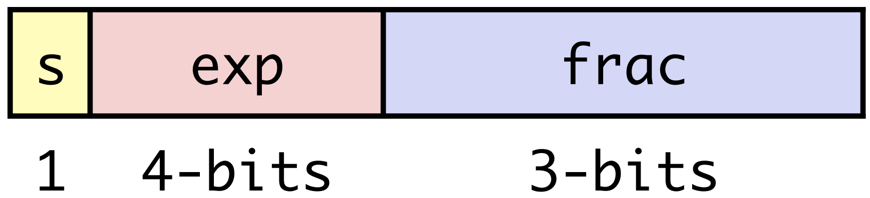
Visualization: Floating Point Encodings



Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point format
- Examples
- Basic conversion
- Properties

Tiny Floating Point Example



- 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`

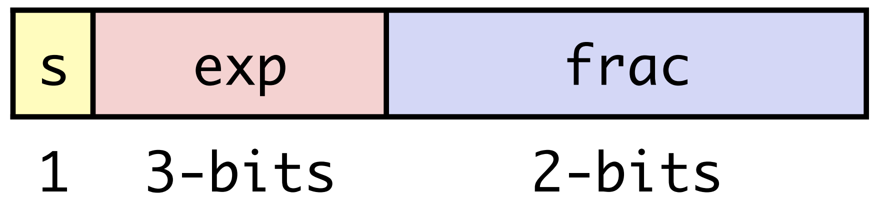
- Same general form as IEEE Format
  - normalized, denormalized
  - representation of 0, NaN, infinity

Dynamic Range (Positive Only)

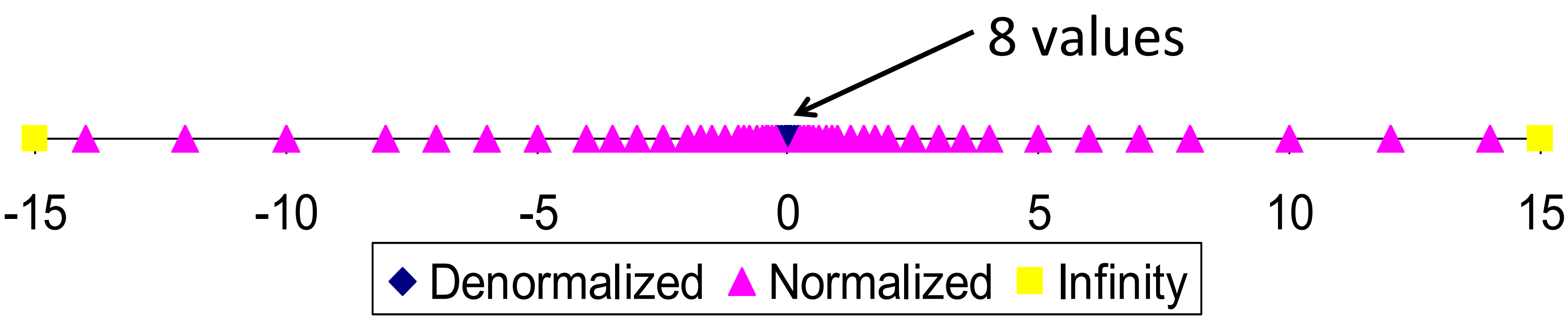
	s	exp	frac	E	Value	
	0	0000	000	-6	0	
	0	0000	001	-6	1/8*1/64 = 1/512	closest to zero
Denormalized numbers	0	0000	010	-6	2/8*1/64 = 2/512	
	…					
	0	0000	110	-6	6/8*1/64 = 6/512	
	0	0000	111	-6	7/8*1/64 = 7/512	largest denorm
	0	0001	000	-6	8/8*1/64 = 8/512	smallest norm
	0	0001	001	-6	9/8*1/64 = 9/512	
	…					
	0	0110	110	-1	14/8*1/2 = 14/16	
	0	0110	111	-1	15/8*1/2 = 15/16	closest to 1 below
Normalized numbers	0	0111	000	0	8/8*1 = 1	
	0	0111	001	0	9/8*1 = 9/8	closest to 1 above
	0	0111	010	0	10/8*1 = 10/8	
	…					
	0	1110	110	7	14/8*128 = 224	
	0	1110	111	7	15/8*128 = 240	largest norm
	0	1111	000	n/a	inf	

Distribution of Values

- 6-bit IEEE-like format
  - e = 3 exponent bits
  - f = 2 fraction bits
  - Bias is 23-1-1 = 3

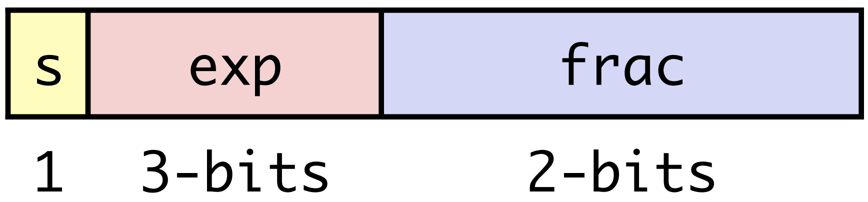


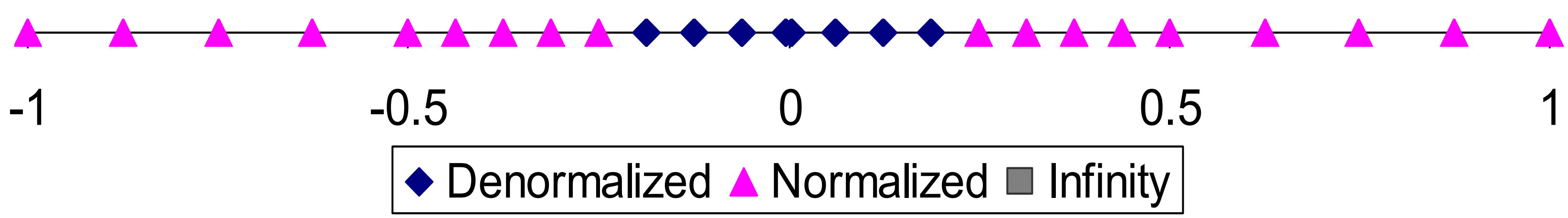
- Notice how the distribution gets denser toward zero.



Distribution of Values (close-up view)

- 6-bit IEEE-like format
  - e = 3 exponent bits
  - f = 2 fraction bits
  - Bias is 3





Interesting Numbers

`{single,double}`

Description	*exp*	*frac*	*Numeric Value*
■ Zero	00…00	00…00	0.0
■ Smallest Pos. Denorm.	00…00	00…01	$2^{-\{23,52\}} \times 2^{-\{126,1022\}}$
▪ Single ≈ 1.4×10^{-45}			
▪ Double ≈ 4.9×10^{-324}			
■ Largest Denormalized	00…00	11…11	$(1.0 - \varepsilon) \times 2^{-\{126,1022\}}$
▪ Single ≈ 1.18×10^{-38}			
▪ Double ≈ 2.2×10^{-308}			
■ Smallest Pos. Normalized	00…01	00…00	$1.0 \times 2^{-\{126,1022\}}$
▪ Just larger than largest denormalized			
■ One	01…11	00…00	1.0
■ Largest Normalized	11…10	11…11	$(2.0 - \varepsilon) \times 2^{\{127,1023\}}$
▪ Single ≈ 3.4×10^{38}			
▪ Double ≈ 1.8×10^{308}			

Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point format
- Examples
- Basic conversion
- Properties

Conversion

- Decimal to float
  - Write binary form
  - Normalize (if possible)
  - Write fractional part
    - Round (covered tomorrow)
  - Compute exponent
    - May be biased (normalized)
    - May be denormalized
  - Write sign bit

- Helpful single-precision values
  - Bias : 127
  - Bits : s=1, exp=8, frac=23

Conversion

- Float to decimal
  - Compute exponent
    - Normalized
    - Denormalized
  - Normalize fractional part (if needed)
  - Compute fractional part
    - Write in binary
  - Adjust binary point
  - Convert to decimal
  - Write sign

Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point format
- Examples
- Basic conversion
- Properties

Special Properties of Encoding

- FP Zero Same as Integer Zero
  - All bits = 0
- Can (Almost) Use Unsigned Integer Comparison
  - Must first compare sign bits
  - Must consider −0 = 0
  - NaNs problematic
    - Will be greater than any other values
    - What should comparison yield?
  - Otherwise OK
    - Denorm vs. normalized
    - Normalized vs. infinity

Summary

- IEEE Floating Point has clear mathematical properties
- Represents numbers of form M x 2^E
- One can reason about operations independent of implementation
  - As if computed with perfect precision and then rounded
- Not the same as real arithmetic
  - Violates associativity/distributivity
  - Makes life difficult for compilers & serious numerical applications programmers