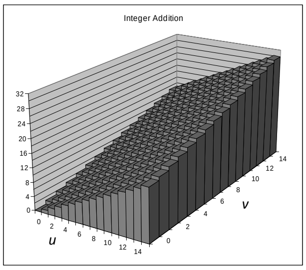
Computer Systems I

Class 3

Visualizing (Mathematical) Integer Addition

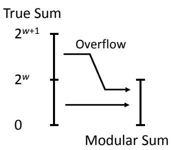
- **Integer Addition**
  - 4-bit integers *u*, *v*
  - Compute true sum $\text{Add}_4(u, v)$
  - Values increase linearly with *u* and *v*
  - Forms planar surface

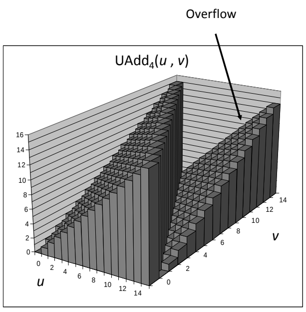
$\text{Add}_4(u, v)$



Visualizing Unsigned Addition

- **Wraps Around**
  - If true sum $\geq 2^w$
  - At most once



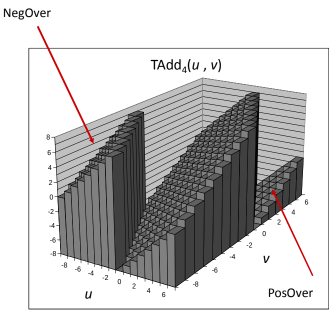


Unsigned Addition

Assuming 8-bit unsigned ints, what would the computed value of 128 + 129 be?

Visualizing 2's Complement Addition

- **Values**
  - 4-bit two's comp.
  - Range from -8 to +7
- **Wraps Around**
  - If sum $\geq 2^{w-1}$
    - Becomes negative
    - At most once
  - If sum $< -2^{w-1}$
    - Becomes positive
    - At most once



Characterizing TAdd

Assuming 8-bit ints, what would the computed value of -127 + -125 be?

Multiplication

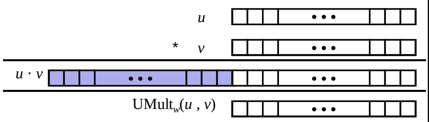
- **Computing Exact Product of *w*-bit numbers *x*, *y***
  - Either signed or unsigned
- **Ranges**
  - Unsigned: $0 \le x * y \le (2^w - 1)^2 = 2^{2w} - 2^{w+1} + 1$
    - Up to $2w$ bits
  - Two's complement min: $x * y \ge (-2^{w-1})*(2^{w-1}-1) = -2^{2w-2} + 2^{w-1}$
    - Up to $2w-1$ bits
  - Two's complement max: $x * y \le (-2^{w-1})^2 = 2^{2w-2}$
    - Up to $2w$ bits, but only for $(TMin_w)^2$ (because of sign bit)
- **Maintaining Exact Results**
  - Would need to keep expanding word size with each product computed
  - Done in software by "arbitrary precision" arithmetic packages

Unsigned Multiplication in C

Operands: *w* bits

True Product: 2**w* bits

Discard *w* bits: *w* bits



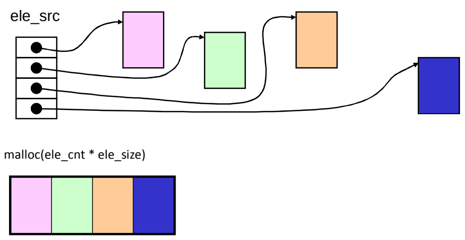
- **Standard Multiplication Function**
  - Ignores high order *w* bits
- **Implements Modular Arithmetic**

  $\mathrm{UMult}_w(u, v) = u \cdot v \bmod 2^w$

Code Security Example #2

- **SUN XDR library**
  - Widely used library for transferring data between machines

```
void* copy_elements(void *ele_src[], int ele_cnt, size_t ele_size);
```



XDR Code

```
void* copy_elements(void *ele_src[], int ele_cnt, size_t ele_size) {
    /*
     * Allocate buffer for ele_cnt objects, each of ele_size bytes
     * and copy from locations designated by ele_src
     */
    void *result = malloc(ele_cnt * ele_size);
    if (result == NULL)
       /* malloc failed */
       return NULL;
    void *next = result;
    int i;
    for (i = 0; i < ele_cnt; i++) {
        /* Copy object i to destination */
        memcpy(next, ele_src[i], ele_size);
       /* Move pointer to next memory region */
       next += ele_size;
    }
    return result;
}
```


XDR Vulnerability

malloc(ele_cnt * ele_size)

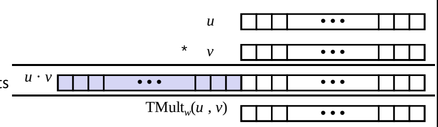
- **What if:**
  - **`ele_cnt`** $= 2^{20} + 1$
  - **`ele_size`** $= 4096$ $= 2^{12}$
  - Allocation = ?? $\boxed{2^{32} + 2^{12} = 4096}$
- **How can I make this function secure?**

Signed Multiplication in C

Operands: *w* bits

True Product: 2**w* bits

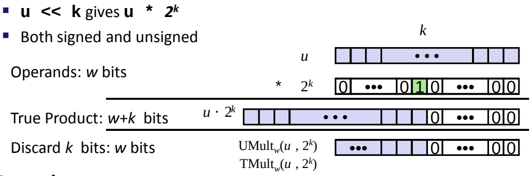
Discard *w* bits: *w* bits



- **Standard Multiplication Function**
  - Ignores high order *w* bits
  - Some of which are different for signed vs. unsigned multiplication
  - Lower bits are the same

Power-of-2 Multiply with Shift

- **Operation**
  - **`u << k`** gives **`u * 2`**^{k}
  - Both signed and unsigned



- **Examples**
  - **`u << 3 == u * 8`**
  - **`u << 5 - u << 3 == u * 24`**
  - Most machines shift and add faster than multiply
    - Compiler generates this code automatically

Compiled Multiplication Code

C Function

```
int mul12(int x)
{
  return x*12;
}
```

Compiled Arithmetic Operations

```
leal (%eax,%eax,2), %eax
sall $2, %eax
```

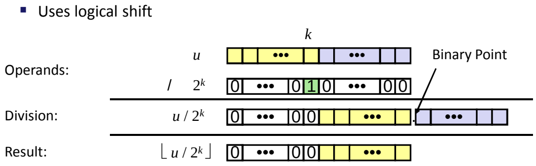
Explanation

```
t <- x+x*2
return t << 2;
```

- **C compiler automatically generates shift/add code when multiplying by constant**

Unsigned Power-of-2 Divide with Shift

- **Quotient of Unsigned by Power of 2**
  - **`u >> k`** gives $\lfloor$ **`u / 2`**^{k} $\rfloor$
  - Uses logical shift



	Division	Computed	Hex	Binary
x	15213	15213	3B 6D	00111011 01101101
x >> 1	7606.5	7606	1D B6	00011101 10110110
x >> 4	950.8125	950	03 B6	00000011 10110110
x >> 8	59.4257813	59	00 3B	00000000 00111011

Compiled Unsigned Division Code

C Function

```
unsigned udiv8(unsigned x)
{
  return x/8;
}
```

Compiled Arithmetic Operations

```
  shrl $3, %eax
```

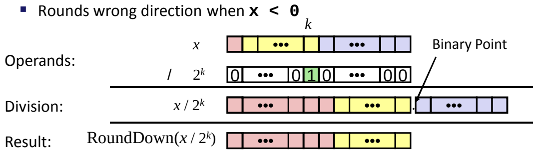
Explanation

```
  # Logical shift
  return x >> 3;
```

- **Uses logical shift for unsigned**
- **For Java Users**
  - Logical shift written as >>>

Signed Power-of-2 Divide with Shift

- **Quotient of Signed by Power of 2**
  - **x >> k** gives ⌊ **x / 2^k** ⌋
  - Uses arithmetic shift
  - Rounds wrong direction when **x < 0**

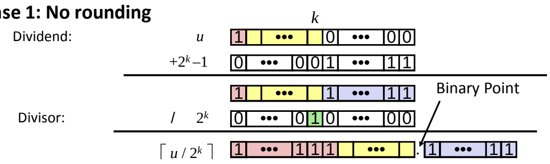


	Division	Computed	Hex	Binary
y	-15213	-15213	C4 93	11000100 10010011
y >> 1	-7606.5	-7607	E2 49	11100010 01001001
y >> 4	-950.8125	-951	FC 49	11111100 01001001
y >> 8	-59.4257813	-60	FF C4	11111111 11000100

Correct Power-of-2 Divide

- **Quotient of Negative Number by Power of 2**
  - Want ⌈ **x / 2^k** ⌉ (Round Toward 0)
  - Compute as ⌊ **(x+2^k-1)/ 2^k** ⌋
    - In C: **(x + (1<<k)-1) >> k**
    - Biases dividend toward 0

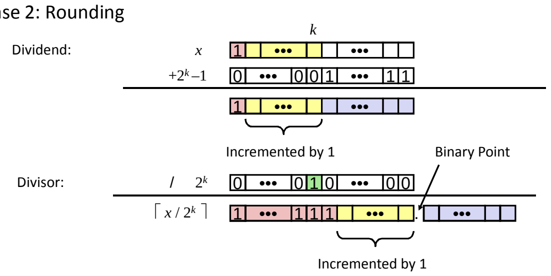
Case 1: No rounding



Biasing has no effect

Correct Power-of-2 Divide (Cont.)

Case 2: Rounding



Biasing adds 1 to final result

Compiled Signed Division Code

C Function

```
int idiv8(int x)
{
  return x/8;
}
```

Compiled Arithmetic Operations

```
  testl %eax, %eax
  js    L4
L3:
  sarl $3, %eax
  ret
L4:
  addl $7, %eax
  jmp  L3
```

Explanation

```
  if x < 0
    x += 7;
  # Arithmetic shift
  return x >> 3;
```

- **Uses arithmetic shift for int**
- **For Java Users**
  - Arith. shift written as >>

Arithmetic: Basic Rules

- **Addition:**
  - Unsigned/signed: Normal addition followed by truncate, same operation on bit level
  - Unsigned: addition mod 2^w
    - Mathematical addition + possible subtraction of 2w
  - Signed: modified addition mod 2^w (result in proper range)
    - Mathematical addition + possible addition or subtraction of 2w

- **Multiplication:**
  - Unsigned/signed: Normal multiplication followed by truncate, same operation on bit level
  - Unsigned: multiplication mod 2^w
  - Signed: modified multiplication mod 2^w (result in proper range)

Arithmetic: Basic Rules

- **Casting between unsigned and signed ints does not change the bits, only the interpretation.**
- **Left shift**
  - Unsigned/signed: multiplication by 2^k
  - Always logical shift
- **Right shift**
  - Unsigned: logical shift, div (division + round to zero) by 2^k
  - Signed: arithmetic shift
    - Positive numbers: div (division + round to zero) by 2^k
    - Negative numbers: div (division + round away from zero) by 2^k
      Use biasing to fix

Integers:

- Representation: unsigned and signed
- Conversion, casting
- Expanding, truncating
- Addition, negation, multiplication, shifting
- **Summary**

Properties of Unsigned Arithmetic

- **Unsigned Multiplication with Addition Forms Commutative Ring**
  - Addition is commutative and associative
  - Multiplication is commutative and associative
  - Multiplication distributes over addtion
    $\text{UMult}_w(t, \text{UAdd}_w(u, v)) = \text{UAdd}_w(\text{UMult}_w(t, u), \text{UMult}_w(t, v))$

Properties of Two's Comp. Arithmetic

- **Comparison to (Mathematical) Integer Arithmetic**
  - Addition and Multiplication are commutative and associative for both
  - Integers obey ordering properties, e.g.,

    $u > 0 \quad \Rightarrow \quad u + v > v$

    $u > 0, v > 0 \quad \Rightarrow \quad u \cdot v > 0$
  - These properties are not obeyed by two's comp. arithmetic

    TMax `+ 1 ==` *TMin*

    `15213 * 30426 == -10030` (16-bit words)

Why Should I Use Unsigned?

- ***Don't*** **Use Just Because Number Nonnegative**
  - Easy to make mistakes

```
unsigned i;
for (i = cnt-2; i >= 0; i--)
  a[i] += a[i+1];
```

  - Can be very subtle

```
#define DELTA sizeof(int)
int i;
for (i = CNT; i-DELTA >= 0; i-= DELTA)
  . . .
```

- ***Do*** **Use When Performing Modular Arithmetic**
  - Multiprecision arithmetic
- ***Do*** **Use When Interested In The Bit Pattern.**

C Puzzle Answers

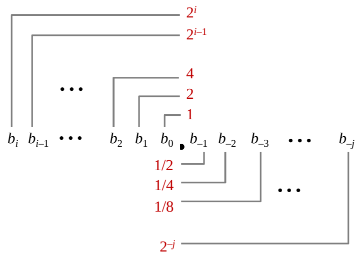
- Assume machine with 32 bit word size, two's comp. integers
- *TMin* makes a good counterexample in many cases

`x < 0`	⇒	`((x*2) < 0)`	False:	*TMin*
`ux >= 0`			True:	0 = *UMin*
`x & 7 == 7`	⇒	`(x<<30) < 0`	True:	$x_1 = 1$
`ux > -1`			False:	0
`x > y`	⇒	`-x < -y`	False:	-1, *TMin*
`x * x >= 0`			False:	30426
`x > 0 && y > 0`	⇒	`x + y > 0`	False:	*TMax*, *TMax*
`x >= 0`	⇒	`-x <= 0`	True:	–*TMax* < 0
`x <= 0`	⇒	`-x >= 0`	False:	*TMin*

Fractional binary numbers

- What is 1011.101?

Fractional Binary Numbers



- **Representation**
  - Bits to right of "binary point" represent fractional powers of 2
  - Represents rational number: $\sum_{k=-j}^{i} b_k \cdot 2^k$

Fractional Binary Numbers: Examples

Value	Representation
5-3/4	101.11_2
2-7/8	10.111_2
63/64	0.111111_2

- **Observations**
  - Divide by 2 by shifting right (binary point shifts left)
  - Multiply by 2 by shifting left (binary point shifts right)
  - Numbers of form $0.111111\ldots_2$ are just below 1.0
    - $1/2 + 1/4 + 1/8 + \ldots + 1/2^i + \ldots \to 1.0$
    - 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**

Value	Representation
1/3	`0.0101010101[01]…`$_2$
1/5	`0.001100110011[0011]…`$_2$
1/10	`0.0001100110011[0011]…`$_2$

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)

Suppose s = 1, E = 011, M = 1.1001.
What is the number?

s	exp	frac

Precisions

- **Single precision: 32 bits**

s	exp	frac
1	8	23

- **Double precision: 64 bits**

s	exp	frac
1	11	52

- **Extended precision: 80 bits (Intel only)**

s	exp	frac
1	15	63 or 64

Normalized Values

- **Condition: `exp` ≠ `000…0` and `exp` ≠ `111…1`**
- **Exponent coded as *biased* value: *E* = *Exp – Bias***
  - *Exp*: unsigned value **`exp`**
  - *Bias* = 2^{e-1} - 1, where e is number of exponent bits
    - This is the value of "TMax for e 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* = $\mathtt{1.\underline{1101101101101}}_2$
  - **`frac`** = $\mathtt{\underline{1101101101101}0000000000}_2$
- **Exponent**
  - *E* = 13
  - *Bias* = 127
  - *Exp* = 140 = $\mathtt{10001100}_2$
- **Result:**

0	10001100	11011011011010000000000
s	exp	frac

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

Floating Point:

- Background: Fractional binary numbers
- IEEE floating point standard: Definition
- **Example and properties**
- Rounding, addition, multiplication
- Floating point in C
- Summary

Tiny Floating Point Example

s	exp	frac
1	4	3

- **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	
Denormalized numbers	0 0000 000	-6	0	
	0 0000 001	-6	1/8*1/64 = 1/512	closest to zero
	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
Normalized numbers	0 0001 000	-6	8/8*1/64 = $\boxed{8/512}$	smallest norm
	…			
	0 1110 110	7	14/8*128 = $\boxed{224}$	
	0 1110 111	7	15/8*128 = 240	largest norm
	0 1111 000	n/a	inf	

What answer should we get when we add these 2 numbers?

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. Normal.	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}			

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

Floating Point:

- Background: Fractional binary numbers
- IEEE floating point standard: Definition
- Example and properties
- **Rounding, addition, multiplication**
- Floating point in C
- Summary

Floating Point Operations: Basic Idea

- `x +f y = Round(x + y)`

- `x ×f y = Round(x × y)`

- **Basic idea**
  - First compute exact result
  - Make it fit into desired precision
    - Possibly overflow if exponent too large
    - Possibly round to fit into `frac`

FP Multiplication

$(-1)^{s1}\ M1\ \ 2^{E1}\ \ \times\ \ (-1)^{s2}\ M2\ \ 2^{E2}$

- **Exact Result:** $(-1)^{s}\ M\ \ 2^{E}$
  - Sign s: $s1$ ^ $s2$
  - Significand M: $M1 * M2$
  - Exponent E: $E1 + E2$

- **Fixing**
  - If $M \geq 2$, shift M right, increment E
  - If E out of range, overflow (or underflow)
  - Round M to fit `frac` precision

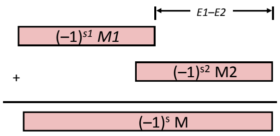
- **Implementation**
  - Biggest chore is multiplying significands

Floating Point Addition

$(-1)^{s1}$ M1 2^{E1} + $(-1)^{s2}$ M2 2^{E2}

Assume $E1 > E2$

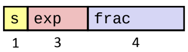
- **Exact Result: $(-1)^s$ M 2^E**
  - Sign s, significand M:
    - Result of signed align & add
  - Exponent E: $E1$



- **Fixing**
  - If $M \geq 2$, shift M right, increment E
  - if $M < 1$, shift M left k positions, decrement E by k
  - Overflow if E out of range
  - Round M to fit `frac` precision

Toy Practice Problem

- **8-bit IEEE-like format**
  - e = 3 exponent bits
  - f = 4 fraction bits
  - Bias is 2^{4-1}-1 = 3 (011)



- **Let's add:**
  - 01110000 + 11110000 + 00010000
  - 01110000 + 00010000 + 11110000

To add 2 numbers:
1. Write them out in full binary fraction form.
2. Add "normally"
3. Convert answer back to IEEE format.
4. Repeat for second addition

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*
    - Overflow and inexactness of rounding
  - 0 is additive identity? *Yes*
  - Every element has additive inverse *Almost*
    - Except for infinities & NaNs
- **Monotonicity**
  - $a \geq b \Rightarrow a+c \geq b+c$? *Almost*
    - Except for infinities & NaNs

Mathematical Properties of FP Mult

- **Compare to Commutative Ring**
  - Closed under multiplication? *Yes*
    - But may generate infinity or NaN
  - Multiplication Commutative? *Yes*
  - Multiplication is Associative? *No*
    - Possibility of overflow, inexactness of rounding
  - 1 is multiplicative identity? *Yes*
  - Multiplication distributes over addition? *No*
    - Possibility of overflow, inexactness of rounding

- **Monotonicity**
  - $a \geq b$ & $c \geq 0 \Rightarrow a * c \geq b * c$? *Almost*
    - Except for infinities & NaNs

Today: Floating Point

- Background: Fractional binary numbers
- IEEE floating point standard: Definition
- Example and properties
- Rounding, addition, multiplication
- **Floating point in C**
- Summary

Floating Point in C

- **C Guarantees Two Levels**

  `float` single precision
  `double` double precision

- **Conversions/Casting**
  - Casting between `int`, `float`, and `double` changes bit representation
  - `Double/float` → `int`
    - Truncates fractional part
    - Like rounding toward zero
    - Not defined when out of range or NaN: Generally sets to TMin
  - `int` → `double`
    - Exact conversion, as long as int has ≤ 53 bit word size
  - `int` → `float`
    - Will round according to rounding mode

Systems I

Answers to Floating Point Puzzles

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

Assume neither d nor f is NAN

• `x == (int)(float) x`	No: 24 bit mantissa
• `x == (int)(double) x`	Yes: 53 bit mantissa
• `f == (float)(double) f`	Yes: increases precision
• `d == (float) d`	No: loses precision
• `f == -(-f);`	Yes: Just change sign bit
• `2/3 == 2/3.0`	No: 2/3 == 0
• `d < 0.0 ⇒ ((d*2) < 0.0)`	Yes!
• `d > f ⇒ -f > -d`	Yes!
• `d * d >= 0.0`	Yes!
• `(d+f)-d == f`	No: Not associative

Systems I

Summary

- **IEEE Floating Point has clear mathematical properties**
- **Represents numbers of form $M \times 2^E$**
- **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