

Bits, Bytes, and Integers

CSE251: System Programming
3rd Lecture, Mar 6

Instructor:
Hyungon Moon

Administrivia

■ Lab 0

- Due tonight (11:59 pm)
- Pre-grading by 1pm today. (25 7/7, 5 6/7, ... etc)
 - Will have short comments about why you lost the score.

■ Lab 1

- Out today, due 3/14 Thu
- Recitation: 3/13 Wed 7:00pm – (voluntarily, location: TBD)

■ Make up class

- 3/14 Thu 7:00pm – 8:15pm (location: TBD)

■ VPN

- Please apply after the class

■ Uni server

- Will reset for you after the class

Today: Bits, Bytes, and Integers

- Representing information as bits
- Bit-level manipulations
- **Integers**
 - Representation: unsigned and signed
 - Conversion, casting
 - Expanding, truncating
 - **Addition, negation, multiplication, shifting**
- Representations in memory, pointers, strings
- Summary

Unsigned Addition

Operands: w bits

True Sum: $w+1$ bits

Discard Carry: w bits

$$\begin{array}{r}
 u \quad \boxed{} \boxed{} \boxed{} | \quad \cdots \quad \boxed{} \boxed{} \boxed{} \\
 + v \quad \boxed{} \boxed{} \boxed{} | \quad \cdots \quad \boxed{} \boxed{} \boxed{} \\
 \hline
 u + v \quad \boxed{\color{red}{1}} \boxed{} \boxed{} \boxed{} | \quad \cdots \quad \boxed{} \boxed{} \boxed{} \\
 \hline
 w(u, v) \quad \boxed{} \boxed{} \boxed{} | \quad \cdots \quad \boxed{} \boxed{} \boxed{}
 \end{array}$$

■ Standard Addition Function

- Ignores carry output

■ Implements Modular Arithmetic

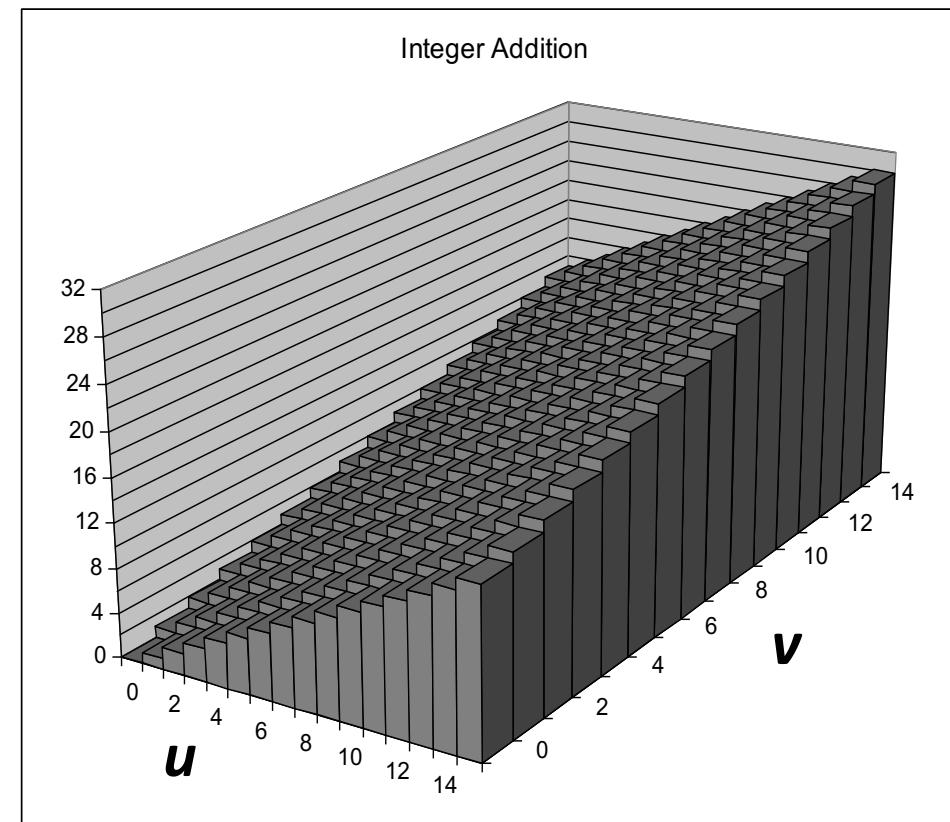
$$s = \text{UAdd}_w(u, v) = u + v \bmod 2^w$$

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

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

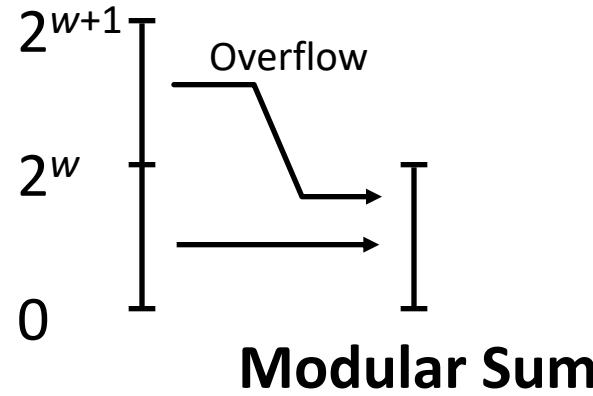


Visualizing Unsigned Addition

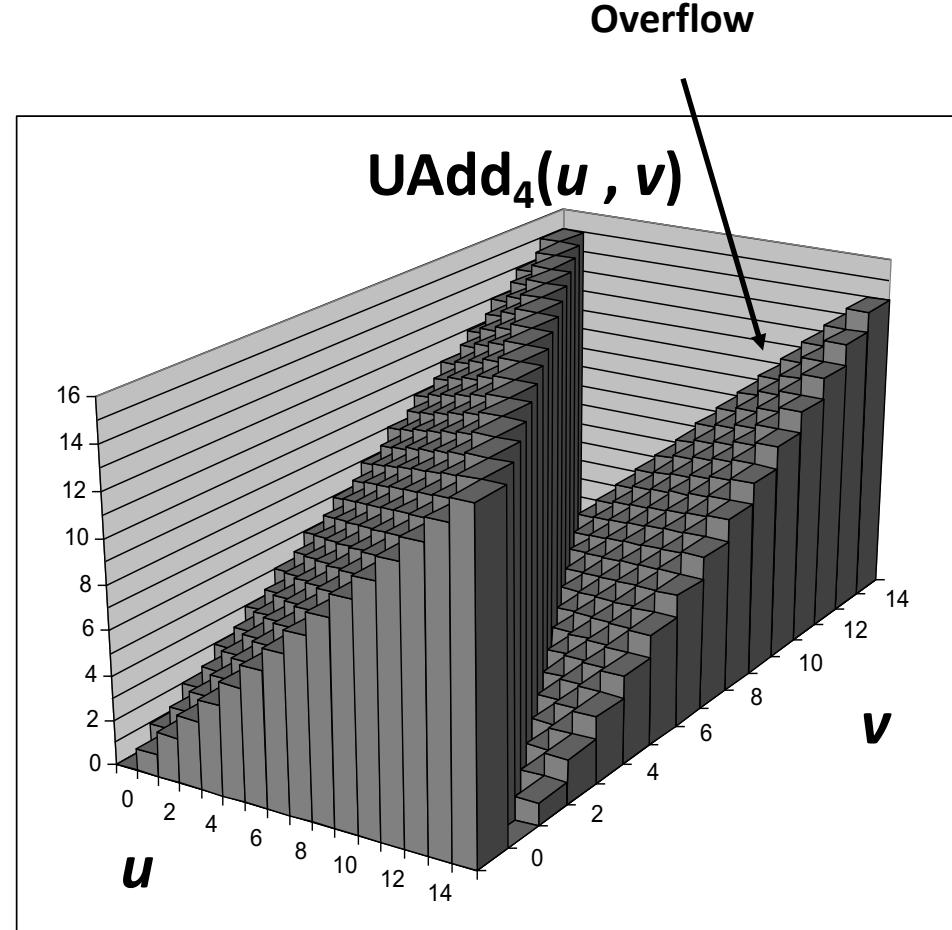
Wraps Around

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

True Sum

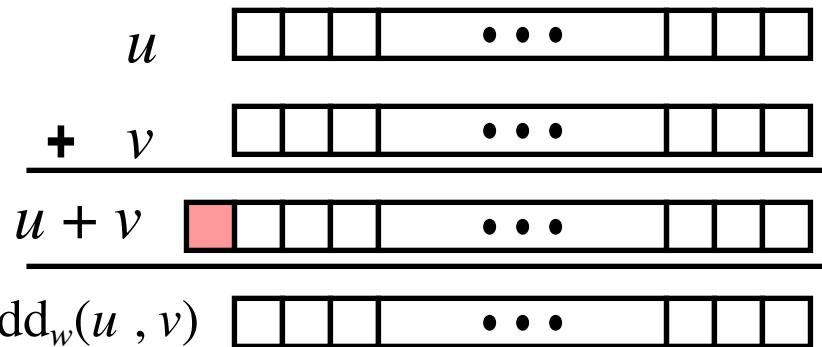


Overflow



Two's Complement Addition

Operands: w bits



True Sum: $w+1$ bits

Discard Carry: w bits

TAdd_w(u , v)

■ TAdd and UAdd have Identical Bit-Level Behavior

- Signed vs. unsigned addition in C:

```
int s, t, u, v;  
s = (int) ((unsigned) u + (unsigned) v);  
t = u + v
```

- Will give $s == t$

TAdd Overflow

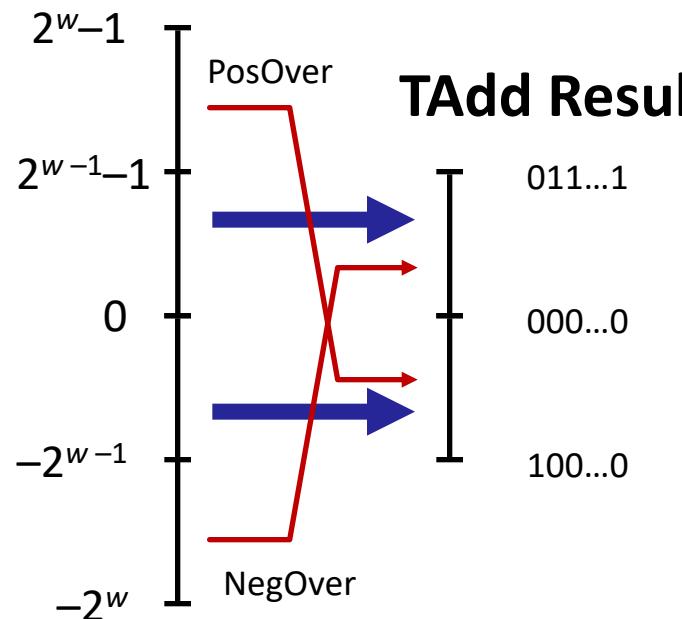
■ Functionality

- True sum requires $w+1$ bits
- Drop off MSB
- Treat remaining bits as 2's comp. integer

0 111...1
0 100...0
0 000...0

1 011...1
1 000...0

True Sum



TAdd Result

Visualizing 2's Complement Addition

■ Values

- 4-bit two's comp.
- Range from -8 to +7

■ Wraps Around

- If $\text{sum} \geq 2^{w-1}$
 - Becomes negative
 - At most once
- If $\text{sum} < -2^{w-1}$
 - Becomes positive
 - At most once

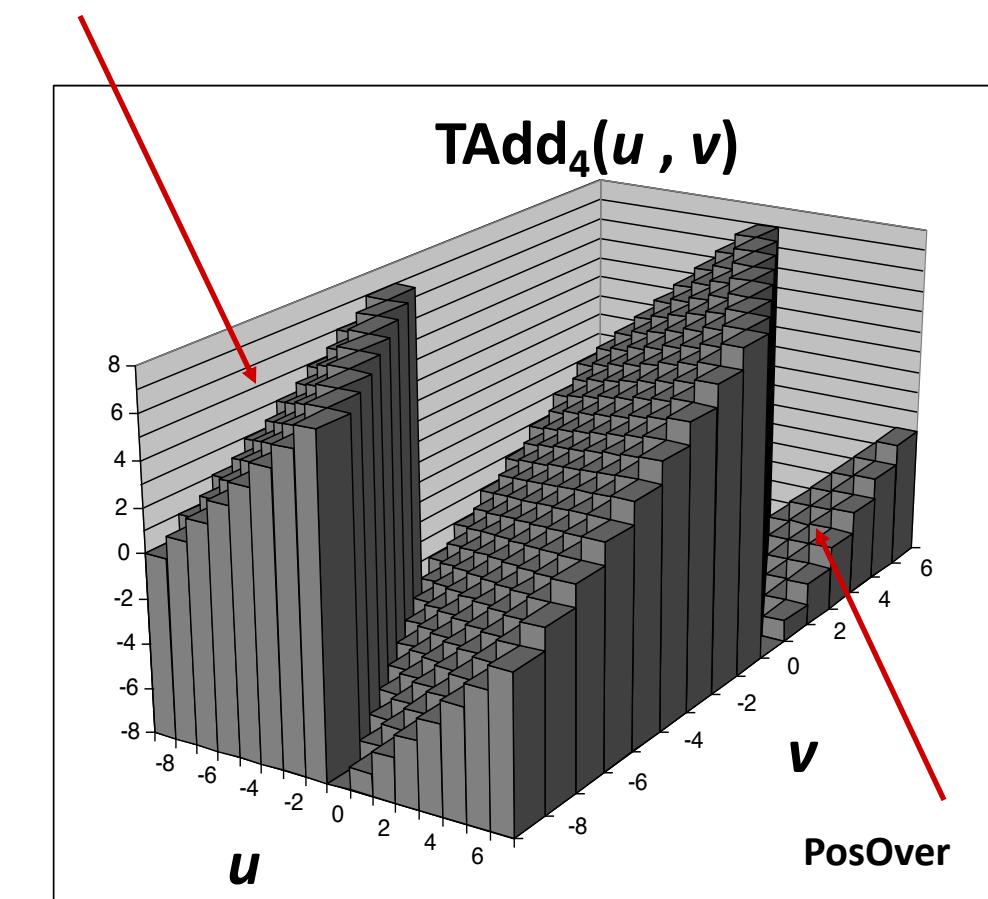
NegOver

TAdd₄(u , v)

u

PosOver

v



Multiplication

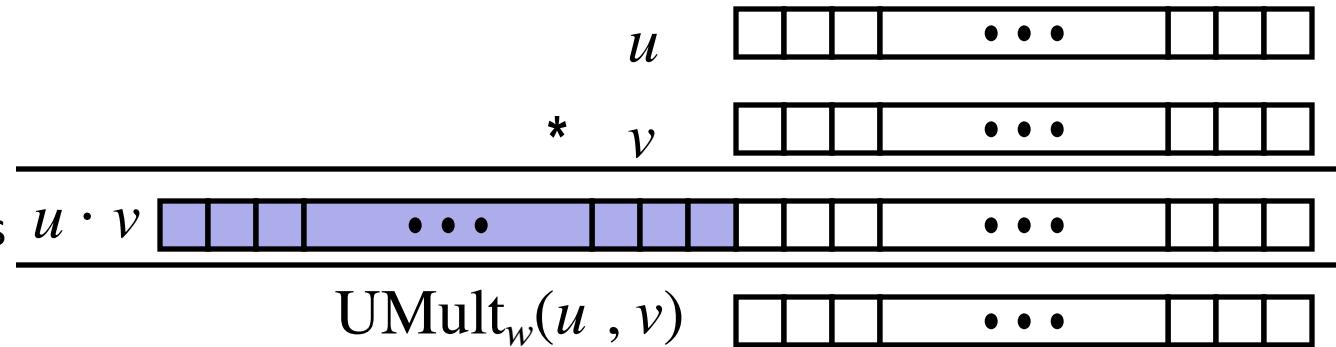
- **Goal: Computing Product of w -bit numbers x, y**
 - Either signed or unsigned
- **But, exact results can be bigger than w bits**
 - Unsigned: up to $2w$ bits
 - Result range: $0 \leq x * y \leq (2^w - 1)^2 = 2^{2w} - 2^{w+1} + 1$
 - Two's complement min (negative): Up to $2w-1$ bits
 - Result range: $x * y \geq (-2^{w-1}) * (2^{w-1} - 1) = -2^{2w-2} + 2^{w-1}$
 - Two's complement max (positive): Up to $2w$ bits, but only for $(TMin_w)^2$
 - Result range: $x * y \leq (-2^{w-1})^2 = 2^{2w-2}$
- **So, maintaining exact results...**
 - would need to keep expanding word size with each product computed
 - is done in software, if needed
 - e.g., 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

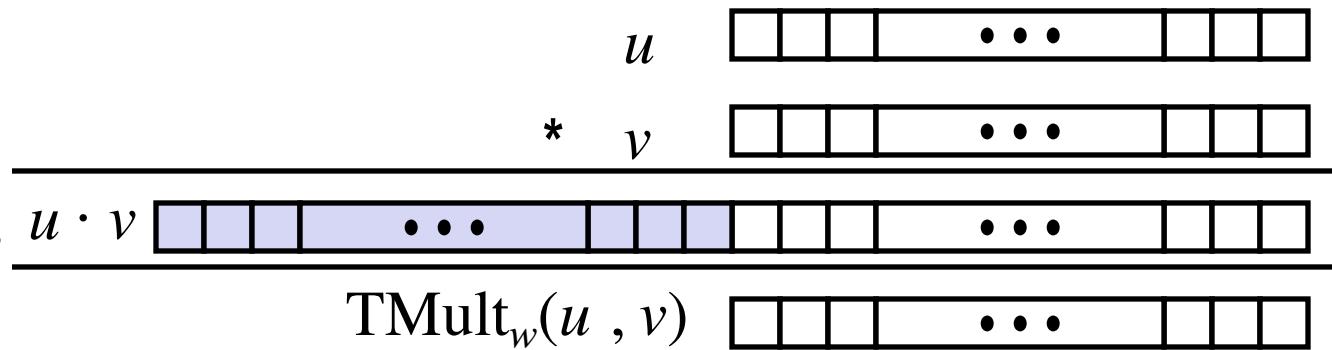
$$UMult_w(u, v) = u \cdot v \bmod 2^w$$

Signed Multiplication in C

Operands: w bits

True Product: $2w$ 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

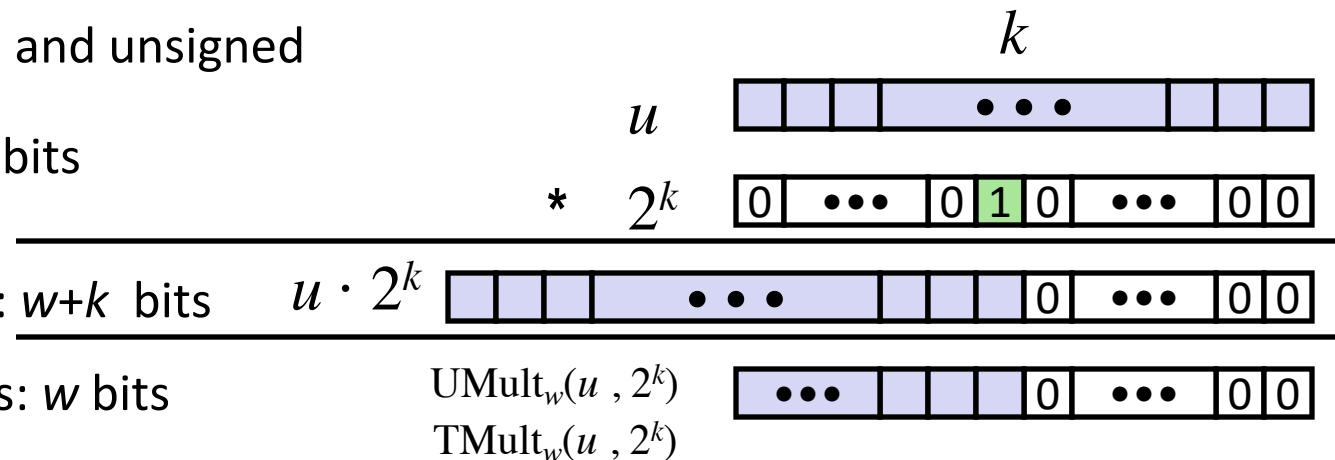
■ Operation

- $u \ll k$ gives $u * 2^k$
- Both signed and unsigned

Operands: w bits

True Product: $w+k$ bits

Discard k bits: w bits



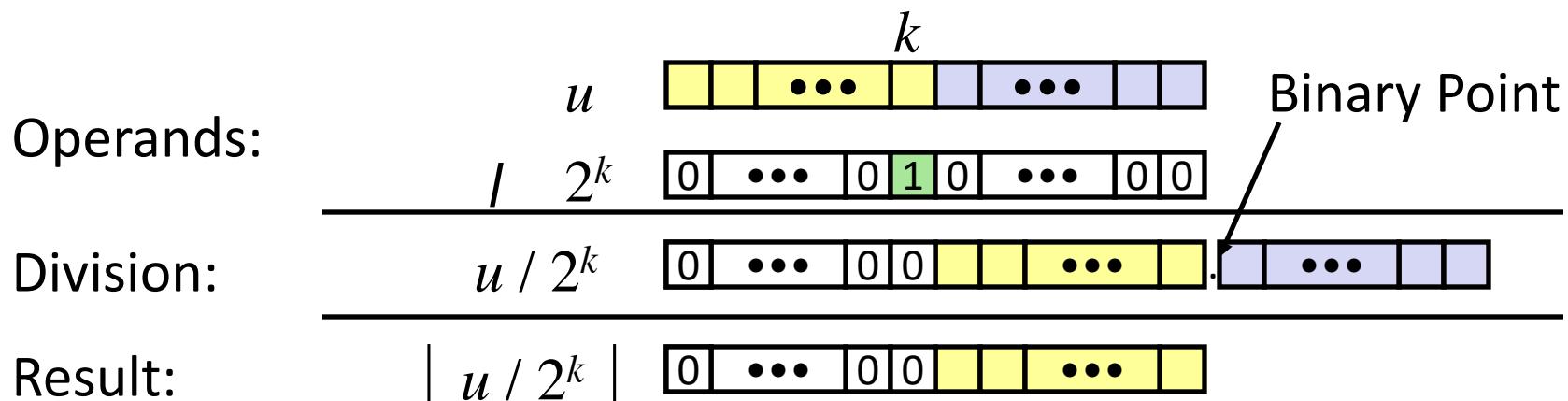
■ Examples

- $u \ll 3 == u * 8$
- $(u \ll 5) - (u \ll 3) == u * 24$
- Most machines shift and add faster than multiply
 - Compiler generates this code automatically

Unsigned Power-of-2 Divide with Shift

■ Quotient of Unsigned by Power of 2

- $u \gg k$ 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 |

Today: Bits, Bytes, and Integers

- Representing information as bits
- Bit-level manipulations
- **Integers**
 - Representation: unsigned and signed
 - Conversion, casting
 - Expanding, truncating
 - Addition, negation, multiplication, shifting
 - **Summary**
- Representations in memory, pointers, strings

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 2^w
- Signed: modified addition mod 2^w (result in proper range)
 - Mathematical addition + possible addition or subtraction of 2^w

■ 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)

Why Should I Use Unsigned?

■ *Don't use without understanding implications*

- 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)  
    . . .
```

Counting Down with Unsigned

■ Proper way to use unsigned as loop index

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

■ See Robert Seacord, *Secure Coding in C and C++*

- C Standard guarantees that unsigned addition will behave like modular arithmetic
 - $0 - 1 \rightarrow UMax$

■ Even better

```
size_t i;  
for (i = cnt-2; i < cnt; i--)  
    a[i] += a[i+1];
```

- Data type `size_t` defined as unsigned value with length = word size
- Code will work even if `cnt = UMax`
- What if `cnt` is signed and < 0 ?

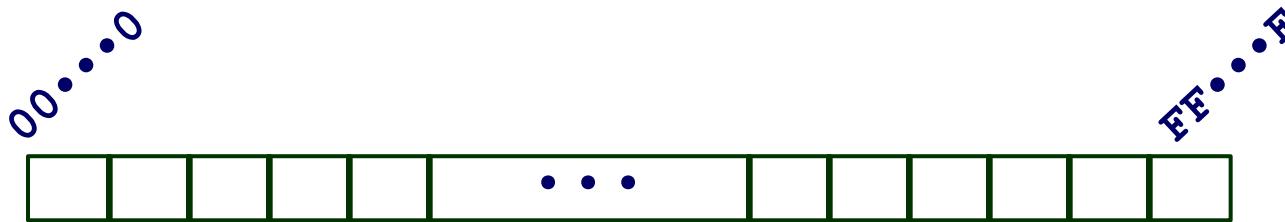
Why Should I Use Unsigned? (cont.)

- ***Do Use When Performing Modular Arithmetic***
 - Multiprecision arithmetic
- ***Do Use When Using Bits to Represent Sets***
 - Logical right shift, no sign extension

Today: Bits, Bytes, and Integers

- Representing information as bits
- Bit-level manipulations
- Integers
 - Representation: unsigned and signed
 - Conversion, casting
 - Expanding, truncating
 - Addition, negation, multiplication, shifting
 - Summary
- Representations in memory, pointers, strings

Byte-Oriented Memory Organization



- **Programs refer to data by address**
 - Conceptually, envision it as a very large array of bytes
 - In reality, it's not, but can think of it that way
 - An address is like an index into that array
 - and, a pointer variable stores an address
- **Note: system provides private address spaces to each “process”**
 - Think of a process as a program being executed
 - So, a program can clobber its own data, but not that of others

Machine Words

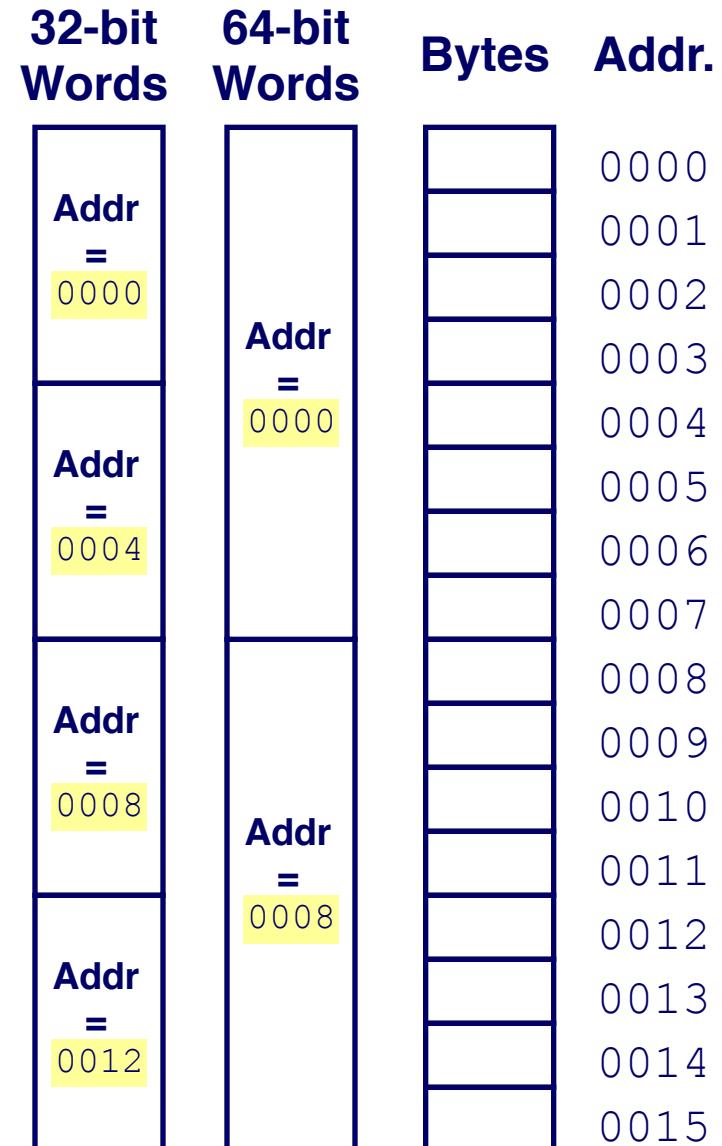
■ Any given computer has a “Word Size”

- Nominal size of integer-valued data
 - and of addresses
- Until recently, most machines used 32 bits (4 bytes) as word size
 - Limits addresses to 4GB (2^{32} bytes)
- Increasingly, machines have 64-bit word size
 - Potentially, could have 18 EB (exabytes) of addressable memory
 - That's 18.4×10^{18}
- Machines still support multiple data formats
 - Fractions or multiples of word size
 - Always integral number of bytes

Word-Oriented Memory Organization

■ Addresses Specify Byte Locations

- Address of first byte in word
- Addresses of successive words differ by 4 (32-bit) or 8 (64-bit)



Example Data Representations

| C Data Type | Typical 32-bit | Typical 64-bit | x86-64 |
|--------------------|----------------|----------------|--------|
| char | 1 | 1 | 1 |
| short | 2 | 2 | 2 |
| int | 4 | 4 | 4 |
| long | 4 | 8 | 8 |
| float | 4 | 4 | 4 |
| double | 8 | 8 | 8 |
| long double | - | - | 10/16 |
| pointer | 4 | 8 | 8 |

Byte Ordering

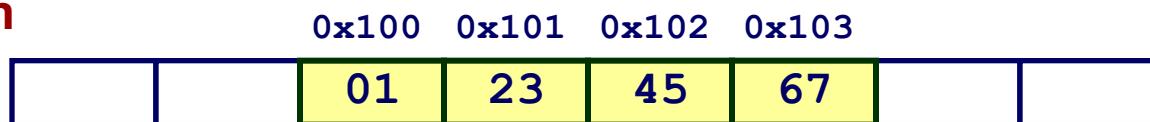
- So, how are the bytes within a multi-byte word ordered in memory?
- Conventions
 - Big Endian: Sun, PPC Mac, Internet
 - Least significant byte has highest address
 - Little Endian: x86, ARM processors running Android, iOS, and Windows
 - Least significant byte has lowest address

Byte Ordering Example

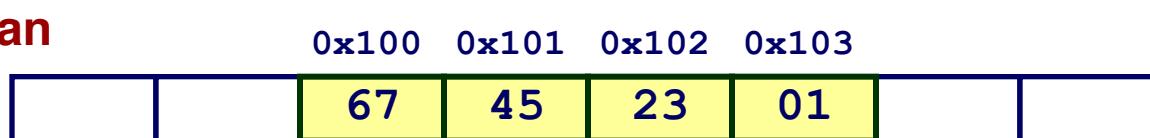
■ Example

- Variable x has 4-byte value of 0x01234567
- Address given by &x is 0x100

BigEndian



Little Endian



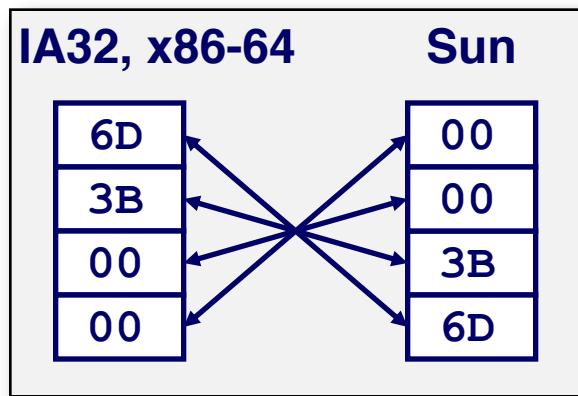
Representing Integers

Decimal: 15213

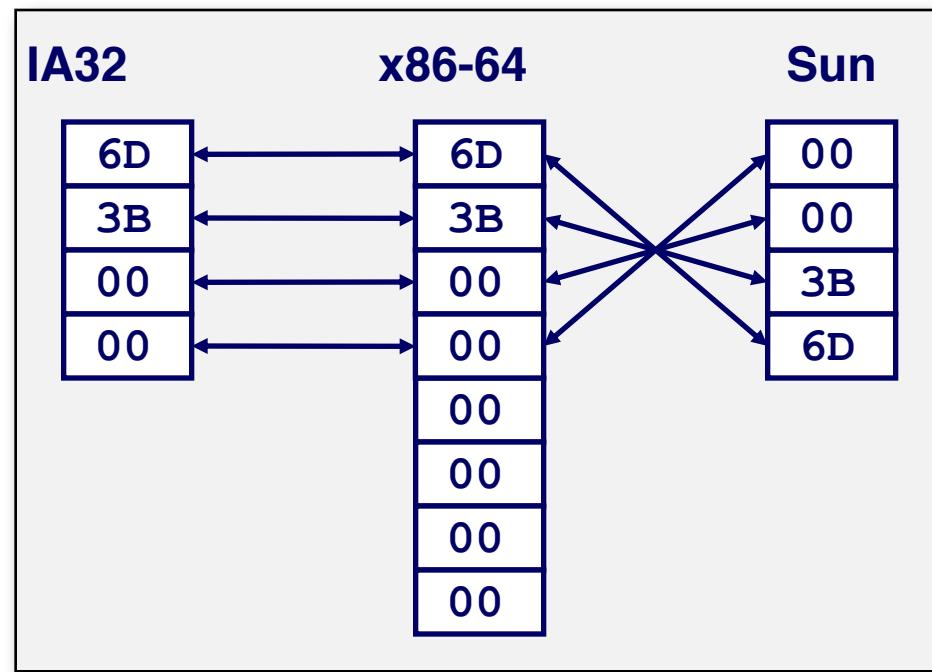
Binary: 0011 1011 0110 1101

Hex: 3 B 6 D

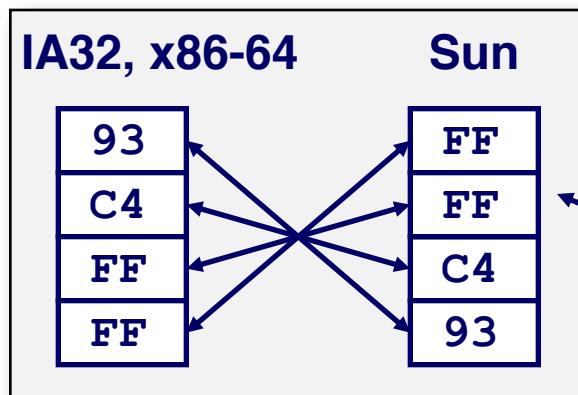
```
int A = 15213;
```



```
long int C = 15213;
```



```
int B = -15213;
```



Two's complement representation

Examining Data Representations

■ Code to Print Byte Representation of Data

- Casting pointer to unsigned char * allows treatment as a byte array

```
typedef unsigned char *pointer;

void show_bytes(pointer start, size_t len){
    size_t i;
    for (i = 0; i < len; i++)
        printf("%p\t0x%.2x\n", start+i, start[i]);
    printf("\n");
}
```

Printf directives:

%p: Print pointer
%x: Print Hexadecimal

show_bytes Execution Example

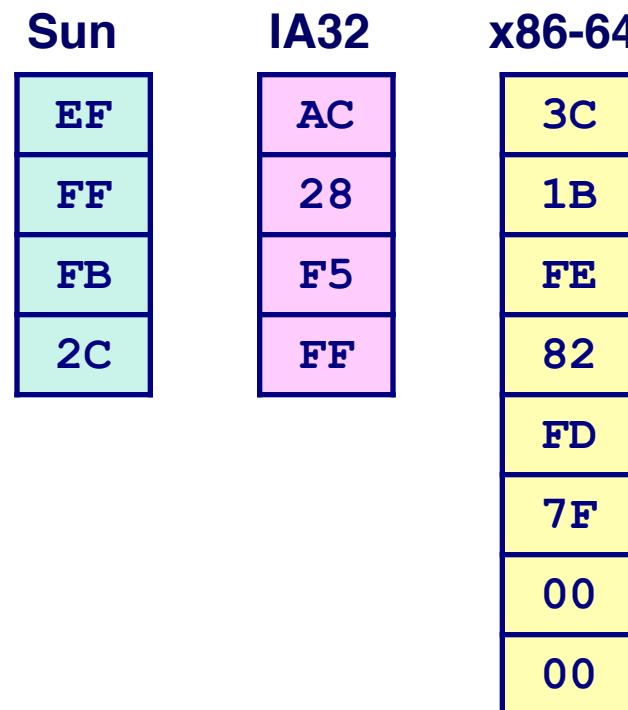
```
int a = 15213;  
printf("int a = 15213;\n");  
show_bytes((pointer) &a, sizeof(int));
```

Result (Linux x86-64):

```
int a = 15213;  
0x7ffb7f71dbc      6d  
0x7ffb7f71dbd      3b  
0x7ffb7f71dbe      00  
0x7ffb7f71dbf      00
```

Representing Pointers

```
int B = -15213;  
int *P = &B;
```



Different compilers & machines assign different locations to objects

Even get different results each time run program

Representing Strings

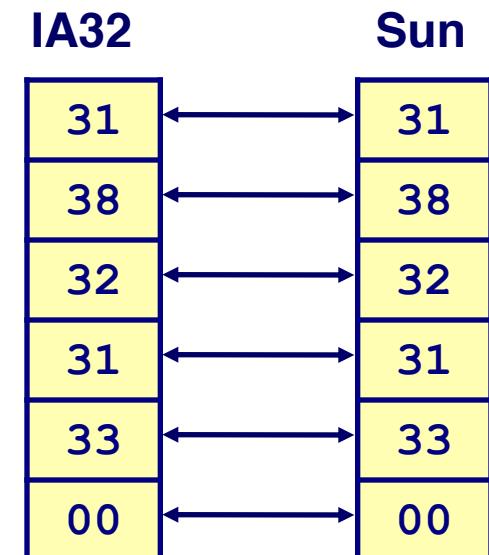
```
char S[6] = "18213";
```

■ Strings in C

- Represented by array of characters
- Each character encoded in ASCII format
 - Standard 7-bit encoding of character set
 - Character “0” has code 0x30
 - Digit i has code $0x30+i$
- String should be null-terminated
 - Final character = 0

■ Compatibility

- Byte ordering not an issue



Integer C Puzzles

Initialization

```
int x = foo();  
int y = bar();  
unsigned ux = x;  
unsigned uy = y;
```

- $x < 0 \rightarrow ((x*2) < 0)$
- $ux \geq 0$
- $x \& 7 == 7 \rightarrow (x<<30) < 0$
- $ux > -1$
- $x > y \rightarrow -x < -y$
- $x * x \geq 0$
- $x > 0 \&& y > 0 \rightarrow x + y > 0$
- $x \geq 0 \rightarrow -x \leq 0$
- $x \leq 0 \rightarrow -x \geq 0$
- $(x|-x)>>31 == -1$
- $ux >> 3 == ux/8$
- $x >> 3 == x/8$
- $x \& (x-1) != 0$

Code Security Example

```
/* Kernel memory region holding user-accessible data */
#define KSIZE 1024
char kbuf[KSIZE];

/* Copy at most maxlen bytes from kernel region to user buffer */
int copy_from_kernel(void *user_dest, int maxlen) {
    /* Byte count len is minimum of buffer size and maxlen */
    int len = KSIZE < maxlen ? KSIZE : maxlen;
    memcpy(user_dest, kbuf, len);
    return len;
}
```

- Similar to code found in FreeBSD's implementation of `getpeername`
- There are legions of smart people trying to find vulnerabilities in programs

Typical Usage

```
/* Kernel memory region holding user-accessible data */
#define KSIZE 1024
char kbuf[KSIZE];

/* Copy at most maxlen bytes from kernel region to user buffer */
int copy_from_kernel(void *user_dest, int maxlen) {
    /* Byte count len is minimum of buffer size and maxlen */
    int len = KSIZE < maxlen ? KSIZE : maxlen;
    memcpy(user_dest, kbuf, len);
    return len;
}
```

```
#define MSIZE 528

void getstuff() {
    char mybuf[MSIZE];
    copy_from_kernel(mybuf, MSIZE);
    printf("%s\n", mybuf);
}
```

Malicious Usage

```
/* Declaration of library function memcpy */
void *memcpy(void *dest, void *src, size_t n);
```

```
/* Kernel memory region holding user-accessible data */
#define KSIZE 1024
char kbuf[KSIZE];

/* Copy at most maxlen bytes from kernel region to user buffer */
int copy_from_kernel(void *user_dest, int maxlen) {
    /* Byte count len is minimum of buffer size and maxlen */
    int len = KSIZE < maxlen ? KSIZE : maxlen;
    memcpy(user_dest, kbuf, len);
    return len;
}
```

```
#define MSIZE 528

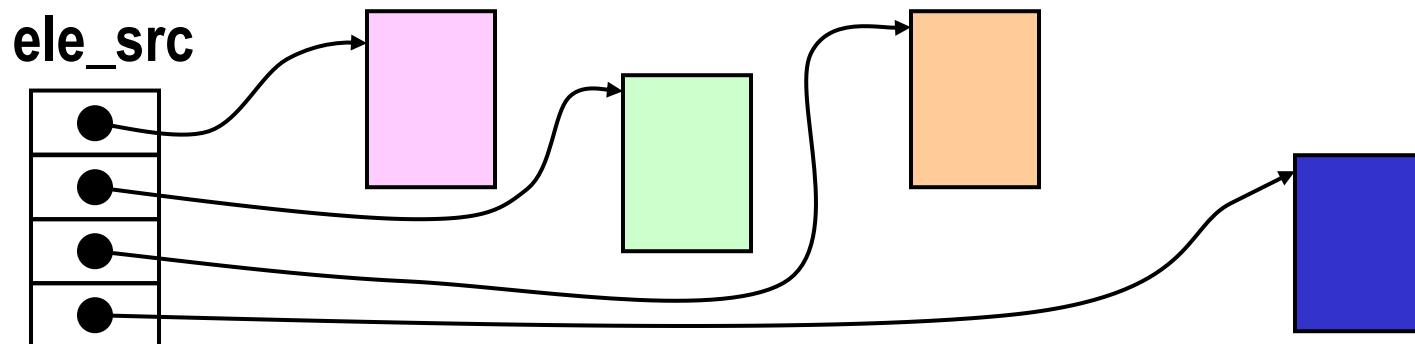
void getstuff() {
    char mybuf[MSIZE];
    copy_from_kernel(mybuf, -MSIZE);
    . . .
}
```

Code Security Example #2

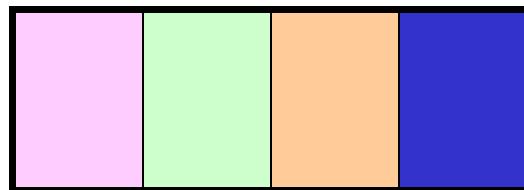
■ SUN XDR library

- Widely used library for transferring data between machines

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



`malloc(ele_cnt * 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 = ??

- How can I make this function secure?

Compiled Multiplication Code

C Function

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

Compiled Arithmetic Operations

```
leaq (%rax,%rax,2), %rax
salq $2, %rax
```

Explanation

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

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

Compiled Unsigned Division Code

C Function

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

Compiled Arithmetic Operations

```
shrq $3, %rax
```

Explanation

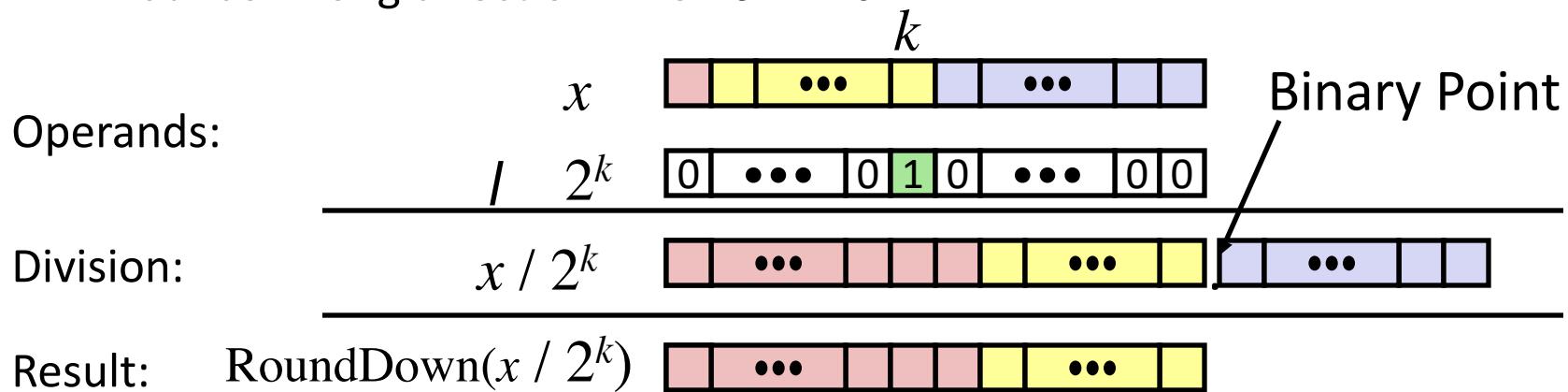
```
# Logical shift
return x >> 3;
```

- 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

- $x \gg k$ gives $\lfloor x / 2^k \rfloor$
- Uses arithmetic shift
- Rounds wrong direction when $u < 0$



| | 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 $\lceil x / 2^k \rceil$ (Round Toward 0) → By definition
 - Compute as $\lfloor (x+2^k-1) / 2^k \rfloor$
 - In C: `(x + (1<<k)-1) >> k`
 - Biases dividend toward 0

Case 1: No rounding

Dividend:

1

14

$$\begin{array}{ccccccc} 1 & \textcolor{yellow}{\boxed{\cdots}} & 0 & \textcolor{yellow}{\boxed{\cdots}} & 0 & 0 \end{array}$$

$$+2^k - 1$$

| | | | | | | | |
|---|-----|---|---|---|-----|---|---|
| 0 | ... | 0 | 0 | 1 | ... | 1 | 1 |
|---|-----|---|---|---|-----|---|---|

Divisor:

$$l = 2^k$$

| | | | | | | | |
|---|-----|---|---|---|-----|---|---|
| 0 | ... | 0 | 1 | 0 | ... | 0 | 0 |
|---|-----|---|---|---|-----|---|---|

$$\lceil u / 2^k \rceil$$

A sequence of colored boxes representing a vector or sequence. The boxes are arranged horizontally. Box 1: Purple, contains the number 1. Box 2: Pink, contains three black dots. Box 3: Purple, contains the numbers 1, 1, 1. Box 4: Yellow. Box 5: Pink, contains three black dots. Box 6: Yellow.

Binary Point

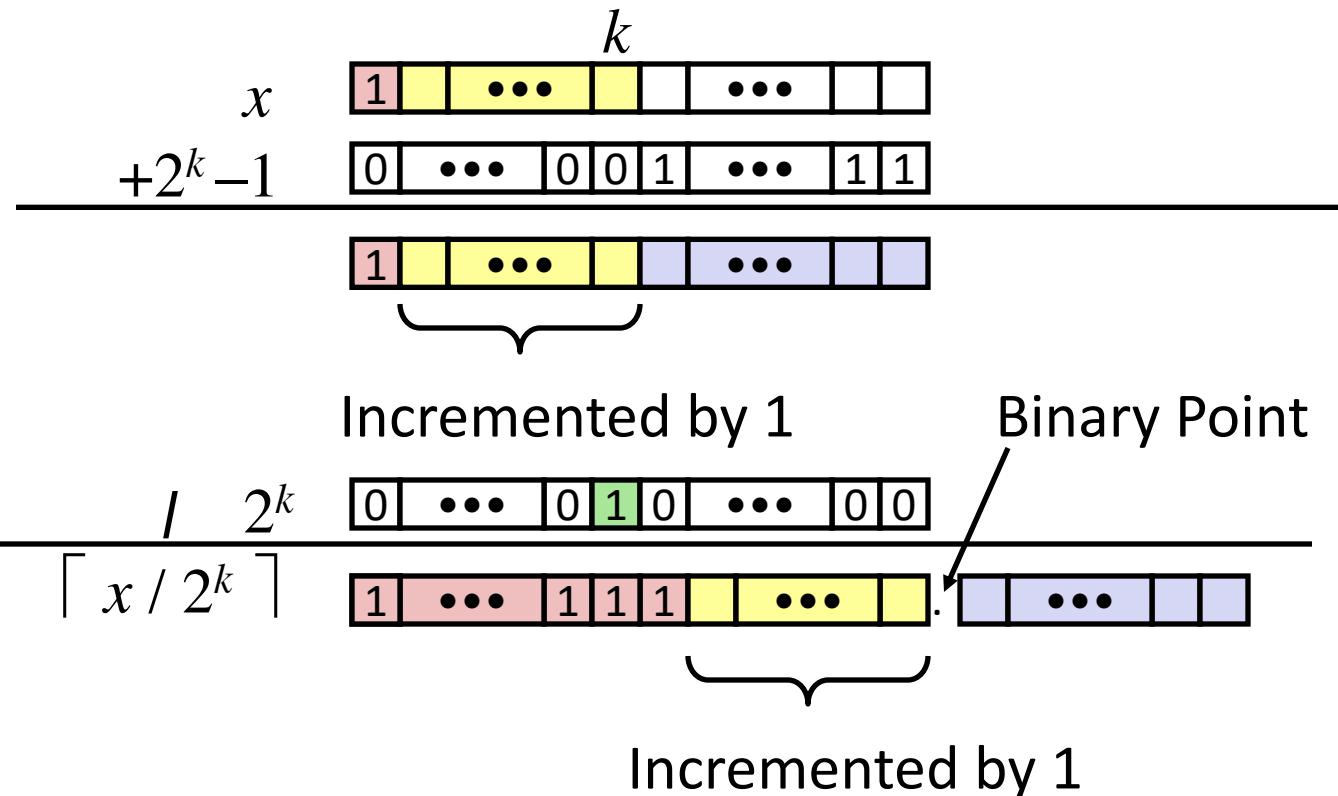
| | | | |
|---|-----|---|---|
| 1 | ... | 1 | 1 |
|---|-----|---|---|

Biasing has no effect

Correct Power-of-2 Divide (Cont.)

Case 2: Rounding

Dividend:



Biasing adds 1 to final result

Compiled Signed Division Code

C Function

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

Compiled Arithmetic Operations

```
testq %rax, %rax
js    L4
L3:
    sarq $3, %rax
    ret
L4:
    addq $7, %rax
    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 >>

Mathematical Properties

■ Modular Addition Forms an *Abelian (commutative) Group*

- **Closed** under addition

$$0 \leq \text{UAdd}_w(u, v) \leq 2^w - 1$$

- **Commutative**

$$\text{UAdd}_w(u, v) = \text{UAdd}_w(v, u)$$

- **Associative: more chance for optimization**

$$\text{UAdd}_w(t, \text{UAdd}_w(u, v)) = \text{UAdd}_w(\text{UAdd}_w(t, u), v)$$

- **0** is additive identity

$$\text{UAdd}_w(u, 0) = u$$

- Every element has additive **inverse**

- Let $\text{UComp}_w(u) = 2^w - u$

$$\text{UAdd}_w(u, \text{UComp}_w(u)) = 0$$

Mathematical Properties of TAdd

■ Isomorphic Group to unsigned with UAdd

- $TAdd_w(u, v) = U2T(UAdd_w(T2U(u), T2U(v)))$
 - Since both have identical bit patterns

■ Two's Complement Under TAdd Forms a Group

- Closed, Commutative, Associative, 0 is additive identity
- Every element has additive inverse

$$TComp_w(u) = \begin{cases} -u & u \neq TMin_w \\ TMin_w & u = TMin_w \end{cases}$$

Negation: Complement & Increment

■ Claim: Following Holds for 2's Complement

$$\sim x + 1 == -x$$

■ Complement

- Observation: $\sim x + x == 1111\dots111 == -1$

$$\begin{array}{r} x \quad \boxed{1} \boxed{0} \boxed{0} \boxed{1} \boxed{1} \boxed{1} \boxed{0} \boxed{1} \\ + \quad \sim x \quad \boxed{0} \boxed{1} \boxed{1} \boxed{0} \boxed{0} \boxed{0} \boxed{1} \boxed{0} \\ \hline -1 \quad \boxed{1} \boxed{1} \boxed{1} \boxed{1} \boxed{1} \boxed{1} \boxed{1} \boxed{1} \end{array}$$

■ Complete Proof?

Arithmetic: Basic Rules

- Unsigned ints, 2's complement ints are isomorphic rings:
isomorphism? (OK to cast-operate-cast)
- 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

Properties of Unsigned Arithmetic

■ Unsigned Multiplication with Addition Forms

Commutative Ring (something like mathematical integer)

- Addition is commutative group
- Closed under multiplication

$$0 \leq \text{UMult}_w(u, v) \leq 2^w - 1$$

- Multiplication Commutative

$$\text{UMult}_w(u, v) = \text{UMult}_w(v, u)$$

- Multiplication is Associative

$$\text{UMult}_w(t, \text{UMult}_w(u, v)) = \text{UMult}_w(\text{UMult}_w(t, u), v)$$

- 1 is multiplicative identity

$$\text{UMult}_w(u, 1) = u$$

- Multiplication distributes over addition

$$\text{UMult}_w(t, \text{UAdd}_w(u, v)) = \text{UAdd}_w(\text{UMult}_w(t, u), \text{UMult}_w(t, v))$$

Properties of Two's Comp. Arithmetic

■ Isomorphic Algebras

- Unsigned multiplication and addition
 - Truncating to w bits
- Two's complement multiplication and addition
 - Truncating to w bits

■ Both Form Rings

- Isomorphic to ring of integers mod 2^w

■ Comparison to (Mathematical) Integer Arithmetic

- Both are rings
- 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)