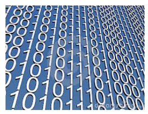
Number Systems and Number Representation

Jennifer Rexford



For Your Amusement

Question: Why do computer programmers confuse Christmas and Halloween?

Answer: Because 25 Dec = 31 Oct

– http://www.electronicsweekly.com

Goals of this Lecture

Help you learn (or refresh your memory) about:

- The binary, hexadecimal, and octal number systems
- Finite representation of unsigned integers
- Finite representation of signed integers
- Finite representation of rational numbers (if time)

Why?

- A power programmer must know number systems and data representation to fully understand C's **primitive data types**

Primitive values and the operations on them

Agenda

Number Systems

Finite representation of unsigned integers

Finite representation of signed integers

Finite representation of rational numbers (if time)

The Decimal Number System

Name

- "decem" (Latin) => ten

Characteristics

- Ten symbols
  - 0 1 2 3 4 5 6 7 8 9
- Positional
  - 2945 ≠ 2495
  - 2945 = $(2*10^3) + (9*10^2) + (4*10^1) + (5*10^0)$

(Most) people use the decimal number system

Why?

The Binary Number System

Name

- "binarius" (Latin) => two

Characteristics

- Two symbols
  - 0 1
- Positional
  - $1010_B \neq 1100_B$

Most (digital) computers use the binary number system

Why?

Terminology

- **Bit**: a binary digit
- **Byte**: (typically) 8 bits

Decimal-Binary Equivalence

Decimal	Binary
0	0
1	1
2	10
3	11
4	100
5	101
6	110
7	111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

Decimal	Binary
16	10000
17	10001
18	10010
19	10011
20	10100
21	10101
22	10110
23	10111
24	11000
25	11001
26	11010
27	11011
28	11100
29	11101
30	11110
31	11111
...	...

7

Decimal-Binary Conversion

Binary to decimal: expand using positional notation

$$\begin{aligned}100101_B &= (1*2^5)+(0*2^4)+(0*2^3)+(1*2^2)+(0*2^1)+(1*2^0)\\ &= 32 + 0 + 0 + 4 + 0 + 1\\ &= 37\end{aligned}$$

8

Decimal-Binary Conversion

Decimal to binary: do the reverse

- Determine largest power of 2 ≤ number; write template

$$37 = (?*2^5)+(?*2^4)+(?*2^3)+(?*2^2)+(?*2^1)+(?*2^0)$$

- Fill in template

$$37 = (1*2^5)+(0*2^4)+(0*2^3)+(1*2^2)+(0*2^1)+(1*2^0)$$

```
 37
-32
  5
 -4
  1
 -1
  0
```

100101_B

9

Decimal-Binary Conversion

Decimal to binary shortcut

- Repeatedly divide by 2, consider remainder

```
37 / 2 = 18 R 1
18 / 2 =  9 R 0
 9 / 2 =  4 R 1
 4 / 2 =  2 R 0
 2 / 2 =  1 R 0
 1 / 2 =  0 R 1
```

Read from bottom to top: 100101_B

10

The Hexadecimal Number System

Name

- "hexa" (Greek) => six
- "decem" (Latin) => ten

Characteristics

- Sixteen symbols
    - 0 1 2 3 4 5 6 7 8 9 A B C D E F
- Positional
    - $A13D_H \neq 3DA1_H$

Computer programmers often use the hexadecimal number system

Why?

11

Decimal-Hexadecimal Equivalence

Decimal	Hex
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	A
11	B
12	C
13	D
14	E
15	F

Decimal	Hex
16	10
17	11
18	12
19	13
20	14
21	15
22	16
23	17
24	18
25	19
26	1A
27	1B
28	1C
29	1D
30	1E
31	1F

Decimal	Hex
32	20
33	21
34	22
35	23
36	24
37	25
38	26
39	27
40	28
41	29
42	2A
43	2B
44	2C
45	2D
46	2E
47	2F
...	...

12

Decimal-Hexadecimal Conversion

Hexadecimal to decimal: expand using positional notation

```
25H = (2*16^1) + (5*16^0)
    =  32     +    5
    =  37
```

Decimal to hexadecimal: use the shortcut

```
37 / 16 = 2 R 5
 2 / 16 = 0 R 2
```

Read from bottom to top: 25_H

13

Binary-Hexadecimal Conversion

Observation: $16^1 = 2^4$

- Every 1 hexadecimal digit corresponds to 4 binary digits

Binary to hexadecimal

```
1010 0001 0011 1101B
  A    1    3    DH
```

Digit count in binary number not a multiple of 4 => pad with zeros on left

Hexadecimal to binary

```
  A    1    3    DH
1010 0001 0011 1101B
```

Discard leading zeros from binary number if appropriate

Is it clear why programmers often use hexadecimal?

14

The Octal Number System

Name

- "octo" (Latin) => eight

Characteristics

- Eight symbols
  - 0 1 2 3 4 5 6 7
- Positional
  - $1743_O \neq 7314_O$

Computer programmers often use the octal number system

Why?

15

Decimal-Octal Equivalence

Decimal	Octal	Decimal	Octal	Decimal	Octal
0	0	16	20	32	40
1	1	17	21	33	41
2	2	18	22	34	42
3	3	19	23	35	43
4	4	20	24	36	44
5	5	21	25	37	45
6	6	22	26	38	46
7	7	23	27	39	47
8	10	24	30	40	50
9	11	25	31	41	51
10	12	26	32	42	52
11	13	27	33	43	53
12	14	28	34	44	54
13	15	29	35	45	55
14	16	30	36	46	56
15	17	31	37	47	57
				...	...

16

Decimal-Octal Conversion

Octal to decimal: expand using positional notation

```
37O = (3*8^1) + (7*8^0)
    =  24    +   7
    =  31
```

Decimal to octal: use the shortcut

```
31 / 8 = 3 R 7
 3 / 8 = 0 R 3
```

Read from bottom to top: 37_O

17

Binary-Octal Conversion

Observation: $8^1 = 2^3$

- Every 1 octal digit corresponds to 3 binary digits

Binary to octal

```
001 010 000 100 111 101B
 1   2   0   4   7   5O
```

Digit count in binary number not a multiple of 3 => pad with zeros on left

Octal to binary

```
 1   2   0   4   7   5O
001 010 000 100 111 101B
```

Discard leading zeros from binary number if appropriate

Is it clear why programmers sometimes use octal?

18

Agenda

Number Systems

Finite representation of unsigned integers

Finite representation of signed integers

Finite representation of rational numbers (if time)

Unsigned Data Types: Java vs. C

Java has type:

- `int`
  - Can represent signed integers

C has type:

- `signed int`
  - Can represent signed integers
- `int`
  - Same as `signed int`
- `unsigned int`
  - Can represent only unsigned integers

To understand C, must consider representation of both unsigned and signed integers

Representing Unsigned Integers

Mathematics

- Range is 0 to ∞

Computer programming

- Range limited by computer's **word** size
- Word size is n bits => range is 0 to $2^n - 1$
- Exceed range => **overflow**

FC010 computers

- n = 64, so range is 0 to $2^{64} - 1$ (huge!)

Pretend computer

- n = 4, so range is 0 to $2^4 - 1$ (15)

Hereafter, assume word size = 4

- All points generalize to word size = 64, word size = n

Representing Unsigned Integers

On pretend computer

Unsigned Integer	Rep
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111
8	1000
9	1001
10	1010
11	1011
12	1100
13	1101
14	1110
15	1111

Adding Unsigned Integers

Addition

```
                1
   3         0011B
+ 10       + 1010B
  --         ----
  13         1101B
```

Start at right column
Proceed leftward
Carry 1 when necessary

```
               11
   7         0111B
+ 10       + 1010B
  --         ----
   1        10001B
```

Beware of overflow

How would you detect overflow programmatically?

Results are mod 2^4

Subtracting Unsigned Integers

Subtraction

```
                12
              0202
  10          1010B
-  7        - 0111B
  --          ----
   3          0011B
```

Start at right column
Proceed leftward
Borrow 2 when necessary

```
              2
   3          0011B
- 10        - 1010B
  --          ----
   9          1001B
```

Beware of overflow

How would you detect overflow programmatically?

Results are mod 2^4

Shifting Unsigned Integers

Bitwise right shift (>> in C): fill on left with zeros

```
10 >> 1 => 5
1010B   0101B

10 >> 2 => 2
1010B   0010B
```

What is the effect arithmetically? (No fair looking ahead)

Bitwise left shift (<< in C): fill on right with zeros

```
5 << 1 => 10
0101B   1010B

3 << 2 => 12
0011B   1100B
```

What is the effect arithmetically? (No fair looking ahead)

Results are mod 2^4

25

Other Operations on Unsigned Ints

Bitwise NOT (~ in C)
- Flip each bit

```
~10 => 5
1010B 0101B
```

Bitwise AND (& in C)
- Logical AND corresponding bits

```
 10       1010B
& 7     & 0111B
 --       ----
  2       0010B
```

Useful for setting selected bits to 0

26

Other Operations on Unsigned Ints

Bitwise OR: (| in C)
- Logical OR corresponding bits

```
 10       1010B
|  1    | 0001B
 --       ----
 11       1011B
```

Useful for setting selected bits to 1

Bitwise exclusive OR (^ in C)
- Logical exclusive OR corresponding bits

```
 10       1010B
^ 10    ^ 1010B
 --       ----
  0       0000B
```

x ^ x sets all bits to 0

27

Aside: Using Bitwise Ops for Arith

Can use <<, >>, and & to do some arithmetic efficiently

x * 2^y == x << y
- 3*4 = 3*2^2 = 3<<2 => 12

Fast way to **multiply** by a power of 2

x / 2^y == x >> y
- 13/4 = 13/2^2 = 13>>2 => 3

Fast way to **divide** by a power of 2

x % 2^y == x & (2^y-1)
- 13%4 = 13%2^2 = 13&(2^2-1) = 13&3 => 1

Fast way to **mod** by a power of 2

```
 13       1101B
& 3     & 0011B
 --       ----
  1       0001B
```

28

Aside: Example C Program

```
#include <stdio.h>
#include <stdlib.h>
int main(void)
{  unsigned int n;
   unsigned int count;
   printf("Enter an unsigned integer: ");
   if (scanf("%u", &n) != 1)
   {  fprintf(stderr, "Error: Expect unsigned int.\n");
      exit(EXIT_FAILURE);
   }
   for (count = 0; n > 0; n = n >> 1)
      count += (n & 1);
   printf("%u\n", count);
   return 0;
}
```

What does it write?

How could this be expressed more succinctly?

29

Agenda

Number Systems

Finite representation of unsigned integers

Finite representation of signed integers

Finite representation of rational numbers (if time)

30

Signed Magnitude

Integer	Rep
-7	1111
-6	1110
-5	1101
-4	1100
-3	1011
-2	1010
-1	1001
-0	1000
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Definition

High-order bit indicates sign

0 => positive

1 => negative

Remaining bits indicate magnitude

$1101_B = -101_B = -5$

$0101_B = 101_B = 5$

31

Signed Magnitude (cont.)

Integer	Rep
-7	1111
-6	1110
-5	1101
-4	1100
-3	1011
-2	1010
-1	1001
-0	1000
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Computing negative

neg(x) = flip high order bit of x

$neg(0101_B) = 1101_B$

$neg(1101_B) = 0101_B$

Pros and cons

+ easy for people to understand

+ symmetric

- two reps of zero

32

Ones' Complement

Integer	Rep
-7	1000
-6	1001
-5	1010
-4	1011
-3	1100
-2	1101
-1	1110
-0	1111
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Definition

High-order bit has weight -7

$1010_B = (1*-7)+(0*4)+(1*2)+(0*1)$
$= -5$

$0010_B = (0*-7)+(0*4)+(1*2)+(0*1)$
$= 2$

33

Ones' Complement (cont.)

Integer	Rep
-7	1000
-6	1001
-5	1010
-4	1011
-3	1100
-2	1101
-1	1110
-0	1111
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Computing negative

neg(x) = ~x

$neg(0101_B) = 1010_B$

$neg(1010_B) = 0101_B$

Computing negative (alternative)

neg(x) = 1111_B - x

$neg(0101_B) = 1111_B - 0101_B$
$= 1010_B$

$neg(1010_B) = 1111_B - 1010_B$
$= 0101_B$

Pros and cons

+ symmetric

- two reps of zero

34

Two's Complement

Integer	Rep
-8	1000
-7	1001
-6	1010
-5	1011
-4	1100
-3	1101
-2	1110
-1	1111
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Definition

High-order bit has weight -8

$1010_B = (1*-8)+(0*4)+(1*2)+(0*1)$
$= -6$

$0010_B = (0*-8)+(0*4)+(1*2)+(0*1)$
$= 2$

35

Two's Complement (cont.)

Integer	Rep
-8	1000
-7	1001
-6	1010
-5	1011
-4	1100
-3	1101
-2	1110
-1	1111
0	0000
1	0001
2	0010
3	0011
4	0100
5	0101
6	0110
7	0111

Computing negative

neg(x) = ~x + 1

neg(x) = onescomp(x) + 1

$neg(0101_B) = 1010_B + 1 = 1011_B$

$neg(1011_B) = 0100_B + 1 = 0101_B$

Pros and cons

- not symmetric

+ one rep of zero

36

Two's Complement (cont.)

Almost all computers use two's complement to represent signed integers

Why?

- Arithmetic is easy
  - Will become clear soon

Hereafter, assume two's complement representation of signed integers

Adding Signed Integers

pos + pos

```
           11
   3     0011B
 + 3   + 0011B
  --     ----
   6     0110B
```

pos + pos (overflow)

```
          111
   7     0111B
 + 1   + 0001B
  --     ----
  -8     1000B
```

pos + neg

```
         1111
   3     0011B
+ -1   + 1111B
  --     ----
   2    10010B
```

How would you detect overflow programmatically?

neg + neg

```
           11
  -3     1101B
+ -2   + 1110B
  --     ----
  -5    11011B
```

neg + neg (overflow)

```
          1 1
  -6     1010B
+ -5   + 1011B
  --     ----
   5    10101B
```


Subtracting Signed Integers

Perform subtraction with borrows — or — Compute two's comp and add

```
          1
         22
   3     0011B
 - 4   - 0100B
  --     ----
  -1     1111B
```

```
   3     0011B
+ -4   + 1100B
  --     ----
  -1     1111B
```

```
  -5     1011B
 - 2   - 0010B
  --     ----
  -7     1001B
```

```
          111
  -5     1011
+ -2   + 1110
  --     ----
  -7    11001
```


Negating Signed Ints: Math

Question: Why does two's comp arithmetic work?

Answer: `[-b] mod 2^4 = [twoscomp(b)] mod 2^4`

```
[-b] mod 2^4
= [2^4 - b] mod 2^4
= [2^4 - 1 - b + 1] mod 2^4
= [(2^4 - 1 - b) + 1] mod 2^4
= [onescomp(b) + 1] mod 2^4
= [twoscomp(b)] mod 2^4
```

See Bryant & O'Hallaron book for much more info

Subtracting Signed Ints: Math

And so:

`[a – b] mod 2^4 = [a + twoscomp(b)] mod 2^4`

```
[a – b] mod 2^4
= [a + 2^4 – b] mod 2^4
= [a + 2^4 – 1 – b + 1] mod 2^4
= [a + (2^4 - 1 – b) + 1] mod 2^4
= [a + onescomp(b) + 1] mod 2^4
= [a + twoscomp(b)] mod 2^4
```

See Bryant & O'Hallaron book for much more info

Shifting Signed Integers

Bitwise left shift (<< in C): fill on right with zeros

```
3 << 1 => 6
0011B   0110B

-3 << 1 => -6
1101B   -1010B
```

What is the effect arithmetically?

Bitwise **arithmetic** right shift: fill on left **with sign bit**

```
6 >> 1 => 3
0110B   0011B

-6 >> 1 => -3
1010B   1101B
```

What is the effect arithmetically?

Results are mod 2^4

Shifting Signed Integers (cont.)

Bitwise **logical** right shift: fill on left **with zeros**

```
6 >> 1 => 3
0110B   0011B

-6 >> 1 => 5
1010B   0101B
```

What is the effect arithmetically???

In C, right shift (>>) could be logical or arithmetic
- Not specified by C90 standard
- Compiler designer decides

Best to avoid shifting signed integers

43

Other Operations on Signed Ints

Bitwise NOT (~ in C)
- Same as with unsigned ints

Bitwise AND (& in C)
- Same as with unsigned ints

Bitwise OR: (| in C)
- Same as with unsigned ints

Bitwise exclusive OR (^ in C)
- Same as with unsigned ints

Best to avoid with signed integers

44

Agenda

Number Systems

Finite representation of unsigned integers

Finite representation of signed integers

Finite representation of rational numbers (if time)

45

Rational Numbers

Mathematics
- A **rational** number is one that can be expressed as the **ratio** of two integers
- Infinite range and precision

Computer science
- Finite range and precision
- Approximate using **floating point** number
  - Binary point "floats" across bits

46

IEEE Floating Point Representation

Common finite representation: **IEEE floating point**
- More precisely: ISO/IEEE 754 standard

Using 32 bits (type float in C):
- 1 bit: sign (0=>positive, 1=>negative)
- 8 bits: exponent + 127
- 23 bits: binary fraction of the form 1.*ddddddddddddddddddddddd*

Using 64 bits (type double in C):
- 1 bit: sign (0=>positive, 1=>negative)
- 11 bits: exponent + 1023
- 52 bits: binary fraction of the form 1.*dddddddddddddddddddddddddddddddddddddddddddddddddddd*

47

Floating Point Example

1 10000011 1011011 0000 000 0000 00000

32-bit representation

Sign (1 bit):
- 1 => negative

Exponent (8 bits):
- 10000011_B = 131
- 131 – 127 = 4

Fraction (23 bits):
- $1.10110110000000000000000_B$
- $1 + (1*2^{-1})+(0*2^{-2})+(1*2^{-3})+(1*2^{-4})+(0*2^{-5})+(1*2^{-6})+(1*2^{-7}) = 1.7109375$

Number:
- $-1.7109375 * 2^4 = -27.375$

48

Floating Point Warning

- Decimal number system can represent only some rational numbers with finite digit count
  - Example: 1/3
- Binary number system can represent only some rational numbers with finite digit count
  - Example: 1/5
- Beware of **roundoff error**
  - Error resulting from inexact representation
  - Can accumulate

Decimal Approx	Rational Value
.3	3/10
.33	33/100
.333	333/1000
...	

Binary Approx	Rational Value
0.0	0/2
0.01	1/4
0.010	2/8
0.0011	3/16
0.00110	6/32
0.001101	13/64
0.0011010	26/128
0.0011001 1	51/256
...	

49

Summary

- The binary, hexadecimal, and octal number systems
- Finite representation of unsigned integers
- Finite representation of signed integers
- Finite representation of rational numbers

- Essential for proper understanding of
  - C primitive data types
  - Assembly language
  - Machine language

50