15-213
“The Class That Gives CMU Its Zip!”

Bits and Bytes
September 2, 2004

Topics

- Why bits?
- Representing information as bits
  - Binary / Hexadecimal
  - Byte representations
    » Numbers
    » Characters and strings
    » Instructions
- Bit-level manipulations
  - Boolean algebra
  - Expressing in C
Why Don’t Computers Use Base 10?

Base 10 Number Representation
- That’s why fingers are known as “digits”
- Natural representation for financial transactions
  - Floating point number cannot exactly represent $1.20$
- Even carries through in scientific notation
  - $15.213 \times 10^3 \quad (1.5213e4)$

Implementing Electronically
- Hard to store
  - ENIAC (First electronic computer) used 10 vacuum tubes / digit
  - IBM 650 used 5+2 bits (1958, successor to IBM’s Personal Automatic Computer, PAC from 1956)
- Hard to transmit
  - Need high precision to encode 10 signal levels on single wire
- Messy to implement digital logic functions
  - Addition, multiplication, etc.
Binary Representations

Base 2 Number Representation

- Represent $15213_{10}$ as $11101101101101_2$
- Represent $1.20_{10}$ as $1.0011001100110011[0011]..._2$
- Represent $1.5213 \times 10^4$ as $1.11011011101101_2 \times 2^{13}$

Electronic Implementation

- Easy to store with bistable elements
- Reliably transmitted on noisy and inaccurate wires
Byte-Oriented Memory Organization

Programs Refer to Virtual Addresses
- Conceptually very large array of bytes
- Actually implemented with hierarchy of different memory types
  - SRAM, DRAM, disk
  - Only allocate for regions actually used by program
- In Unix and Windows NT, address space private to particular “process”
  - Program being executed
  - Program can clobber its own data, but not that of others

Compiler + Run-Time System Control Allocation
- Where different program objects should be stored
- Multiple mechanisms: static, stack, and heap
- In any case, all allocation within single virtual address space
Encoding Byte Values

Byte = 8 bits

- Binary \( 00000000_2 \) to \( 11111111_2 \)
- Decimal: \( 0_{10} \) to \( 255_{10} \)
  - First digit must not be 0 in C
- Octal: \( 000_8 \) to \( 0377_8 \)
  - Use leading 0 in C
- Hexadecimal \( 00_{16} \) to \( FF_{16} \)
  - Base 16 number representation
  - Use characters ‘0’ to ‘9’ and ‘A’ to ‘F’
  - Write FA1D37B_{16} in C as \( 0x\text{FA1D37B} \)
  - Or \( 0xf\text{a1d37b} \)

<table>
<thead>
<tr>
<th>Hex</th>
<th>Decimal</th>
<th>Binary</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
<td>0000</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0001</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>0010</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>0011</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0100</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>0101</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>0110</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>0111</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>1000</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>1001</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>1010</td>
</tr>
<tr>
<td>B</td>
<td>11</td>
<td>1011</td>
</tr>
<tr>
<td>C</td>
<td>12</td>
<td>1100</td>
</tr>
<tr>
<td>D</td>
<td>13</td>
<td>1101</td>
</tr>
<tr>
<td>E</td>
<td>14</td>
<td>1110</td>
</tr>
<tr>
<td>F</td>
<td>15</td>
<td>1111</td>
</tr>
</tbody>
</table>
Literary Hex

Common 8-byte hex filler:

- 0xdeadbeef
- Can you think of other 8-byte fillers?
Machine Words

Machine Has “Word Size”

- Nominal size of integer-valued data
  - Including addresses
- Most current machines use 32 bits (4 bytes) words
  - Limits addresses to 4GB
  - Becoming too small for memory-intensive applications
- High-end systems use 64 bits (8 bytes) words
  - Potential address space \( \approx 1.8 \times 10^{19} \) bytes
- Machines 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)
# Data Representations

## Sizes of C Objects (in Bytes)

<table>
<thead>
<tr>
<th>C Data Type</th>
<th>Alpha (RIP)</th>
<th>Typical 32-bit</th>
<th>Intel IA32</th>
</tr>
</thead>
<tbody>
<tr>
<td>unsigned</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>int</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>long int</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>char</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>short</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>float</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>double</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>long double</td>
<td>8/16(^{+})</td>
<td>8</td>
<td>10/12</td>
</tr>
<tr>
<td>char *</td>
<td>8</td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

(\(^{+}\): Depends on compiler&OS, 128bit FP is done in software)
Byte Ordering

How should bytes within multi-byte word be ordered in memory?

Conventions

- Sun’s, Mac’s are “Big Endian” machines
  - Least significant byte has highest address
- Alphas, PC’s are “Little Endian” machines
  - Least significant byte has lowest address
# Byte Ordering Example

**Big Endian**
- Least significant byte has highest address

**Little Endian**
- Least significant byte has lowest address

**Example**
- Variable $x$ has 4-byte representation 0x01234567
- Address given by &x is 0x100

<table>
<thead>
<tr>
<th>Big Endian</th>
<th>0x100 0x101 0x102 0x103</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>01 23 45 67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Little Endian</th>
<th>0x100 0x101 0x102 0x103</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>67 45 23 01</td>
</tr>
</tbody>
</table>
Reading Byte-Reversed Listings

Disassembly

- Text representation of binary machine code
- Generated by program that reads the machine code

Example Fragment

<table>
<thead>
<tr>
<th>Address</th>
<th>Instruction Code</th>
<th>Assembly Rendition</th>
</tr>
</thead>
<tbody>
<tr>
<td>8048365:</td>
<td>5b</td>
<td>pop %ebx</td>
</tr>
<tr>
<td>8048366:</td>
<td>81 c3 ab 12 00 00</td>
<td>add $0x12ab,%ebx</td>
</tr>
<tr>
<td>804836c:</td>
<td>83 bb 28 00 00 00 00</td>
<td>cmpl $0x0,0x28 (%ebx)</td>
</tr>
</tbody>
</table>

Deciphering Numbers

- Value: 0x12ab
- Pad to 4 bytes: 0x000012ab
- Split into bytes: 00 00 12 ab
- Reverse: ab 12 00 00
Examining Data Representations

Code to Print Byte Representation of Data

- Casting pointer to unsigned char * creates byte array

```c
typedef unsigned char *pointer;

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

Printf directives:
- %p: Print pointer
- %x: Print Hexadecimal
**show_bytes** Execution Example

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

**Result (Linux):**

```c
int a = 15213;
0x11fffffffcb8  0x6d
0x11fffffffcb9  0x3b
0x11fffffffcba  0x00
0x11fffffffcbbb 0x00
```
Representing Integers

```c
int A = 15213;
int B = -15213;
long int C = 15213;
```

Decimal: 15213
Binary: 0011 1011 0110 1101
Hex: 3 B 6 D

Two’s complement representation (Covered next lecture)
Representing Pointers

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

Alpha Address
Hex:   1 F F F F F F C A 0
Binary: 0001 1111 1111 1111 1111 1111 1100 1010 0000

Sun Address
Hex:   E F F F F F B 2 C
Binary: 1110 1111 1111 1111 1111 1111 1011 0010 1100

Linux Address
Hex:   B F F F F F 8 D 4
Binary: 1011 1111 1111 1111 1111 1000 1101 0100

Different compilers & machines assign different locations to objects
Representing Floats

Float \( F = 15213.0; \)

Linux/Alpha \( F \)  
\[ \begin{array}{c}
00 \\
B4 \\
6D \\
46 \\
\end{array} \]

Sun \( F \)  
\[ \begin{array}{c}
46 \\
6D \\
B4 \\
00 \\
\end{array} \]

IEEE Single Precision Floating Point Representation

<table>
<thead>
<tr>
<th>Hex:</th>
<th>4</th>
<th>6</th>
<th>6</th>
<th>D</th>
<th>B</th>
<th>4</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Binary:</td>
<td>0100</td>
<td>0110</td>
<td>0110</td>
<td>1101</td>
<td>1011</td>
<td>0100</td>
<td>0000</td>
<td>0000</td>
</tr>
</tbody>
</table>

15213: 1110 1101 1011 01

Not same as integer representation, but consistent across machines

Can see some relation to integer representation, but not obvious
Representing Strings

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
- Text files generally platform independent
  - Except for different conventions of line termination character(s)!
    » Unix (‘\n’ = 0x0a = ^J)
    » Mac (‘\r’ = 0x0d = ^M)
    » DOS and HTTP (‘\r\n’ = 0x0d0a = ^M^J)

```cpp
char S[6] = "15213";
```
Machine-Level Code Representation

Encode Program as Sequence of Instructions

- Each simple operation
  - Arithmetic operation
  - Read or write memory
  - Conditional branch
- Instructions encoded as bytes
  - Alpha’s, Sun’s, Mac’s use 4 byte instructions
    » Reduced Instruction Set Computer (RISC)
  - PC’s use variable length instructions
    » Complex Instruction Set Computer (CISC)
- Different instruction types and encodings for different machines
  - Most code is not binary compatible

Programs are Byte Sequences Too!
Representing Instructions

```
int sum(int x, int y)
{
    return x+y;
}
```

- For this example, Alpha & Sun use two 4-byte instructions
  - Use differing numbers of instructions in other cases
- PC uses 7 instructions with lengths 1, 2, and 3 bytes
  - Same for NT and for Linux
  - NT / Linux not fully binary compatible

Different machines use totally different instructions and encodings
**Boolean Algebra**

**Developed by George Boole in 19th Century**

- Algebraic representation of logic
  - Encode “True” as 1 and “False” as 0

<table>
<thead>
<tr>
<th>And</th>
<th>Or</th>
</tr>
</thead>
<tbody>
<tr>
<td>A&amp;B = 1 when both A=1 and B=1</td>
<td>A</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>&amp;</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Not

<table>
<thead>
<tr>
<th>~A = 1 when A=0</th>
</tr>
</thead>
<tbody>
<tr>
<td>~</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>

- Exclusive-Or (Xor)

<table>
<thead>
<tr>
<th>A$\land$B = 1 when either A=1 or B=1, but not both</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\land$</td>
</tr>
<tr>
<td>0</td>
</tr>
<tr>
<td>1</td>
</tr>
</tbody>
</table>
Application of Boolean Algebra

Applied to Digital Systems by Claude Shannon

- 1937 MIT Master’s Thesis
- Reason about networks of relay switches
  - Encode closed switch as 1, open switch as 0

\[ A \& \sim B \]

Connection when

\[ A \& \sim B \mid \sim A \& B \]

= \[ A^\sim B \]
Integer Algebra

Integer Arithmetic

- \langle \mathbb{Z}, +, *, -, 0, 1 \rangle \text{ forms a “ring”}
- Addition is “sum” operation
- Multiplication is “product” operation
- \( - \) is additive inverse
- 0 is identity for sum
- 1 is identity for product
Boolean Algebra

Boolean Algebra

- \( \langle \{0, 1\}, |, &, \sim, 0, 1 \rangle \) forms a “Boolean algebra”
- Or is “sum” operation
- And is “product” operation
- \( \sim \) is “complement” operation (not additive inverse)
- 0 is identity for sum
- 1 is identity for product
Boolean Algebra \iff Integer Ring

- **Commutativity**
  \[
  A \lor B = B \lor A \quad A + B = B + A \\
  A \land B = B \land A \quad A \cdot B = B \cdot A
  \]

- **Associativity**
  \[
  (A \lor B) \lor C = A \lor (B \lor C) \quad (A + B) + C = A + (B + C) \\
  (A \land B) \land C = A \land (B \land C) \quad (A \cdot B) \cdot C = A \cdot (B \cdot C)
  \]

- **Product distributes over sum**
  \[
  A \land (B \lor C) = (A \land B) \lor (A \land C) \\
  A \cdot (B + C) = A \cdot B + A \cdot C
  \]

- **Sum and product identities**
  \[
  A \lor 0 = A \quad A + 0 = A \\
  A \land 1 = A \quad A \cdot 1 = A
  \]

- **Zero is product annihilator**
  \[
  A \land 0 = 0 \quad A \cdot 0 = 0
  \]

- **Cancellation of negation**
  \[
  \sim (\sim A) = A \quad \neg (\neg A) = A
  \]
Boolean Algebra $\neq$ Integer Ring

- **Boolean: Sum distributes over product**
  \[ A \lor (B \land C) = (A \lor B) \land (A \lor C) \quad A + (B \ast C) \neq (A + B) \ast (A + C) \]

- **Boolean: Idempotency**
  \[ A \lor A = A \quad A + A \neq A \]
  - “A is true” or “A is true” = “A is true”
  \[ A \land A = A \quad A \ast A \neq A \]

- **Boolean: Absorption**
  \[ A \lor (A \land B) = A \quad A + (A \ast B) \neq A \]
  - “A is true” or “A is true and B is true” = “A is true”
  \[ A \land (A \lor B) = A \quad A \ast (A + B) \neq A \]

- **Boolean: Laws of Complements**
  \[ A \lor \neg A = 1 \quad A + \neg A \neq 1 \]
  - “A is true” or “A is false”

- **Ring: Every element has additive inverse**
  \[ A \lor \neg A \neq 0 \quad A + \neg A = 0 \]
Boolean Ring

- \(\langle \{0,1\}, ^\wedge, \& , I, 0, 1 \rangle\)
- Identical to integers mod 2
- \(I\) is identity operation: \(I(A) = A\)
  \[ A ^\wedge A = 0 \]

Properties of & and ^

<table>
<thead>
<tr>
<th>Property</th>
<th>Boolean Ring</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commutative sum</td>
<td>(A ^\wedge B = B ^\wedge A)</td>
</tr>
<tr>
<td>Commutative product</td>
<td>(A &amp; B = B &amp; A)</td>
</tr>
<tr>
<td>Associative sum</td>
<td>((A ^\wedge B) ^\wedge C = A ^\wedge (B ^\wedge C))</td>
</tr>
<tr>
<td>Associative product</td>
<td>((A &amp; B) &amp; C = A &amp; (B &amp; C))</td>
</tr>
<tr>
<td>Prod. over sum</td>
<td>(A &amp; (B ^\wedge C) = (A &amp; B) ^\wedge (A &amp; C))</td>
</tr>
<tr>
<td>0 is sum identity</td>
<td>(A ^\wedge 0 = A)</td>
</tr>
<tr>
<td>1 is prod. identity</td>
<td>(A &amp; 1 = A)</td>
</tr>
<tr>
<td>0 is product annihilator</td>
<td>(A &amp; 0 = 0)</td>
</tr>
<tr>
<td>Additive inverse</td>
<td>(A ^\wedge A = 0)</td>
</tr>
</tbody>
</table>
Relations Between Operations

DeMorgan’s Laws

- Express & in terms of |, and vice-versa
  - $A \land B = \neg(\neg A \lor \neg B)$
    - A and B are true if and only if neither A nor B is false
  - $A \lor B = \neg(\neg A \land \neg B)$
    - A or B are true if and only if A and B are not both false

Exclusive-Or using Inclusive Or

- $A \oplus B = (\neg A \land B) \lor (A \land \neg B)$
  - Exactly one of A and B is true
- $A \oplus B = (A \lor B) \land \neg(A \land B)$
  - Either A is true, or B is true, but not both
General Boolean Algebras

Operate on Bit Vectors

- Operations applied bitwise

\[
\begin{array}{cccc}
01101001 & 01101001 & 01101001 \\
\& 01010101 & | 01010101 & ^ 01010101 & \sim 01010101 \\
01000001 & 01111101 & 00111100 & 10101010
\end{array}
\]

All of the Properties of Boolean Algebra Apply
Representing & Manipulating Sets

Representation

- **Width** \( w \) bit vector represents subsets of \( \{0, \ldots, w-1\} \)
- \( a_j = 1 \) if \( j \in A \)
  
  \[
  \begin{array}{c|c}
  01101001 & \{0, 3, 5, 6\} \\
  76543210 & \\
  \end{array}
  \]

  \[
  \begin{array}{c|c}
  01010101 & \{0, 2, 4, 6\} \\
  76543210 & \\
  \end{array}
  \]

Operations

- **&** Intersection  \( 01000001 \) \( \{0, 6\} \)
- **|** Union  \( 01111101 \) \( \{0, 2, 3, 4, 5, 6\} \)
- **^** Symmetric difference  \( 00111100 \) \( \{2, 3, 4, 5\} \)
- **~** Complement  \( 10101010 \) \( \{1, 3, 5, 7\} \)
Bit-Level Operations in C

Operations &, |, ~, ^ Available in C

- Apply to any "integral" data type
  - long, int, short, char, unsigned
- View arguments as bit vectors
- Arguments applied bit-wise

Examples (Char data type)

- \( \sim 0x41 \rightarrow 0xBE \)
  \( \sim 01000001_2 \rightarrow 10111110_2 \)
- \( \sim 0x00 \rightarrow 0xFF \)
  \( \sim 00000000_2 \rightarrow 11111111_2 \)
- \( 0x69 \ & \ 0x55 \rightarrow 0x41 \)
  \( 01101001_2 \ & \ 01010101_2 \rightarrow 01000001_2 \)
- \( 0x69 \ | \ 0x55 \rightarrow 0x7D \)
  \( 01101001_2 \ | \ 01010101_2 \rightarrow 01111101_2 \)
Contrast: Logic Operations in C

Contrast to Logical Operators

- &&, ||, !
  - View 0 as “False”
  - Anything nonzero as “True”
  - Always return 0 or 1
  - Early termination

Examples (char data type)

- !0x41 --> 0x00
- !0x00 --> 0x01
- !!0x41 --> 0x01

- 0x69 && 0x55 --> 0x01
- 0x69 || 0x55 --> 0x01
- p && *p (avoids null pointer access)
Shift Operations

Left Shift: \( x \ll y \)
- Shift bit-vector \( x \) left \( y \) positions
  - Throw away extra bits on left
  - Fill with 0’s on right

Right Shift: \( x \gg y \)
- Shift bit-vector \( x \) right \( y \) positions
  - Throw away extra bits on right
- Logical shift
  - Fill with 0’s on left
- Arithmetic shift
  - Replicate most significant bit on right
  - Useful with two’s complement integer representation

<table>
<thead>
<tr>
<th>Argument ( x )</th>
<th>( 01100010 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( &lt;&lt; 3 )</td>
<td>( 00010000 )</td>
</tr>
<tr>
<td>Log. ( &gt;&gt; 2 )</td>
<td>( 00011000 )</td>
</tr>
<tr>
<td>Arith. ( &gt;&gt; 2 )</td>
<td>( 00011000 )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Argument ( x )</th>
<th>( 10100010 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( &lt;&lt; 3 )</td>
<td>( 00010000 )</td>
</tr>
<tr>
<td>Log. ( &gt;&gt; 2 )</td>
<td>( 00101000 )</td>
</tr>
<tr>
<td>Arith. ( &gt;&gt; 2 )</td>
<td>( 11101000 )</td>
</tr>
</tbody>
</table>
Cool Stuff with Xor

- Bitwise Xor is form of addition
- With extra property that every value is its own additive inverse
  \[ A \oplus A = 0 \]

```c
void funny(int *x, int *y)
{
    *x = *x ^ *y;  /* #1 */
    *y = *x ^ *y;  /* #2 */
    *x = *x ^ *y;  /* #3 */
}
```
More Fun with Bitvectors

Bit-board representation of chess position:

```c
unsigned long long blk_king, wht_king, wht_rook_mv2, ...;
```

```
8   0 1 2
7
6
5
4
3
2   61 62 63
1   a b c d e f g h
```
More Bitvector Magic

Count the number of 1’s in a word

MIT Hackmem 169:

```c
int bitcount(unsigned int n)
{
    unsigned int tmp;

    tmp = n - ((n >> 1) & 0333333333333)
         - ((n >> 2) & 0111111111111;
    return ((tmp + (tmp >> 3)) & 030707070707) % 63;
}
```
Some Other Uses for Bitvectors

Representation of small sets

Representation of polynomials:
- Important for error correcting codes
- Arithmetic over finite fields, say GF(2^n)
- Example 0x15213 : \(x^{16} + x^{14} + x^{12} + x^9 + x^4 + x + 1\)

Representation of graphs
- A ‘1’ represents the presence of an edge

Representation of bitmap images, icons, cursors, ...
- Exclusive-or cursor patent

Representation of Boolean expressions and logic circuits
Summary of the Main Points

It’s All About Bits & Bytes
- Numbers
- Programs
- Text

Different Machines Follow Different Conventions for
- Word size
- Byte ordering
- Representations

Boolean Algebra is the Mathematical Basis
- Basic form encodes “false” as 0, “true” as 1
- General form like bit-level operations in C
  - Good for representing & manipulating sets