Computer Systems
Lecture 5

Overview

- Alphanumeric characters
- ASCII code
- Unicode
- Representation of numbers
- Declaration of variables in programs

Alphanumeric characters

- Data that computers deal with usually comprise:
  - Letters, such as 'A', 'B', etc.
  - Numbers, such as '1', '9', etc.
  - Other characters, such as punctuation marks.

Binary codes of alphanumeric characters

- The need for coding.
  - Since all data are represented by binary sequences in computers, the question comes up as to how they should be represented.
- The choice of code may be arbitrary.
- But, what matters is consistency, i.e. **standards** are needed.
  - Can you see why?

Alphanumeric codes

- **ASCII** code (**A**merican **S**tandard **C**ode for **I**nformation **i**nterchange (**7-bit** code) and its extensions (8-bit codes) (well-established).

- **EBCDIC** code (**E**xtended **B**inary **C**oded **D**ecimal **I**nterchange **C**ode) **8-bit** code. (IBM mainframe computers)

- **Unicode.** Recent **16-bit** standard. (Up to 2^{16} characters can be encoded)

ASCII code table

```
     0    1    2    3    4    5    6    7    8    9    A    B    C    D    E    F
    --------------------------------------------------------------------------------
0|  NUL  SOH  STX  ETX  EOT  ENQ  ACK  BEL  BS   HT   LF   VT   FF   CR   SO   SI
1|  DLE  DC1  DC2  DC3  DC4  NAK  SYN  ETB  CAN  EM        SUB  ESC  FS   GS   RS   US
2|  SP   !    "    #    $    %    &    '    (    )    *    +    ,    -    .    /
3|  0    1    2    3    4    5    6    7    8    9    :    ;    <    =    >    ?
4|  @    A    B    C    D    E    F    G    H    I    J    K    L    M    N    O
5|  P    Q    R    S    T    U    V    W    X    Y    Z    [    \    ]    ^    _
6|  `    a    b    c    d    e    f    g    h    i    j    k    l    m    n    o
7|  p    q    r    s    t    u    v    w    x    y    z    {    |    }    ~    DEL
```

ASCII code table (cont.)

- It is 7-bit code, so the code of any character can be represented as a two-digit hexadecimal number (row's and column's numbers are the first and second digits, respectively).

- 3 bits are represented by the first hexadecimal digit and the next 4-bits are represented by the second hexadecimal digit.

ASCII code table (cont.)

NUL	null
SOH	start of heading
STX	start of text
ETX	end of text
EOT	end of transmission
ENQ	enquiry
ACK	acknowledge
BEL	bell
…	…

- Check http://www.jimprice.com/jim-asc.shtml for more information.

- The following `C` code should sound your bell.

```
#include <stdio.h>
void main(){
   putchar(7);
}
```

ASCII code table (cont.)

- Only half of possible byte (8-bits) patterns is used.

- The second half is used in the 8-bit **extension(s)** to represent the codes of additional symbols, line shapes, foreign letters, etc.

ASCII code table (cont.)

- The order of the letter codes is compatible with the alphabetical order of the letters.

- The table is divided into two classes of codes:
  - Printing characters.
  - Control characters.

ASCII code table (cont.)

- Printing characters produce output on the screen or on a printer.

- Control characters are used:
  - To control the position of output on the screen or paper (e.g. 'HT').
  - To cause some action to occur (e.g. 'BEL').
  - To communicate status between the computer and an I/O device (e.g. 'Control-C' combination).

Limitations of ASCII code

- The limitations of the well-established 8 bit ASCII codes.
  - Too limited for the display requirements of modern Windows-based word-processors.
  - The requirement of global software market for handling **international** character sets.

Unicode

- Even 8-bit extensions of ASCII code table is capable to code only up to 256 characters.

- Unicode Standard (1991) is an 16-bit international encoding system for information interchange.

  – Code values are available for more than 65,000 characters.

Unicode (cont.)

- Standard includes the European alphabetic scripts, Middle Eastern right-to-left scripts, and scripts of Asia, Africa and America, ideographic characters of China, Japan, etc

- Standard includes punctuation marks, mathematical symbols, geometrical shapes, etc.

- For Unicode Chart:
  http://www.ssec.wisc.edu/~tomw/java/unicode.html

Unicode in Java

- Is it OK to use Unicode in Java programming?
  - In a Java program, if you want to show some Chinese characters, you have to use their Unicode representations.

Representation of numbers

- Any representation of numbers capable of supporting arithmetic manipulation has to deal with both **integers** and **real values**, **positive** and **negative**.
  - Integers are whole numbers, with no fraction part.
  - Real numbers extend beyond the decimal point.

Representation of integers

- Generally 4 bytes, i.e. 32 bits.
- Two's complement as a method of representing and manipulating negative integers.

-128	64	32	16	8	4	2	1

$$00000101_2 = 5_{10} \qquad 11111011_2 = -5_{10}$$

- For the advantages of Two's complement representations for integers, you can read:
  - [Wil06] section 5.3. or
  - Wikipedia's article about Two's complement.
    - Put "Two's complement" there and launch a search!

Representation of real numbers

- IEEE 754 standard.
  - The most widely-used standard for floating-point computation.
  - defines formats for representing floating-point numbers, special values, and a set of floating-point operations that operate on these values.
- Check the article on "IEEE 754" in Wikipedia for more information.

Declaration of variables in programs

- What happens when you declare variables in a program?
  - You are telling the compiler to reserve the correct amount of **memory space** to hold the variable.
  - You are also telling the compiler what **encoding/decoding**/representation scheme to be used.

char letter	1 byte
short count	2 bytes
long world_population	4 bytes
long double world_weight	8 bytes

Q&A

- Q: The most widely used binary code with non–IBM mainframes and virtually all microcomputers, is
- a. EBCDIC
- b. DOS
- c. ASCII
- d. LAN

Q&A

- Q. Which of the following coding schemes is not yet commonly used?

- a. Unicode

- b. EBCDIC

- c. ASCII

- d. All of the above are common coding schemes.

Q&A

- Q. A drawback to ASCII is that it
- a. cannot handle all the characters of some languages other than English and European languages.
- b. uses only 4 bits to form each character.
- c. is slower than EBCDIC.
- d. None of the above is correct.

Q&A

- ASCII is divided into two classes of codes?

- What is the most widely-used standard for floating-point computation?

- Two things happen when you declare variables in a program. What are they?

Readings

- [Wil06] Chapter 2, section 2.8 for coding.
- [Wil06] Chapter 5, section 5.3 for Two's complement.
- Wikipedia.
    - http://en.wikipedia.org/wiki/Unicode
    - http://en.wikipedia.org/wiki/Two%27s_complement
    - http://en.wikipedia.org/wiki/IEEE_754