BLM5207
Computer Organization

Prof. Dr. Nizamettin AYDIN
naydin@yildiz.edu.tr
http://www3.yildiz.edu.tr/~naydin

Data Formats

1

1

Data Formats

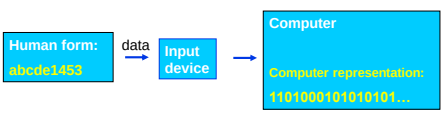
- Computers
  - Process and store all forms of data in binary format
- Human communication
  - Includes language, images and sounds
- Data formats:
  - Specifications for converting data into computer-usable form
  - Define the different ways human data may be represented, stored and processed by a computer

2

2

Sources of Data

- Binary input
  - Begins as discrete input
    - Example: keyboard input such as A 1+2=3 math
      - Keyboard generates a binary number code for each key
- Analog
  - Continuous data such as sound or images
    - Requires hardware to convert data into binary numbers



3

3

Common Data Representations

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC
Image (bitmapped)	▪GIF (graphical image format) ▪TIF (tagged image file format) ▪PNG (portable network graphics)
Image (object)	PostScript, JPEG, SWF (Macromedia Flash), SVG
Outline graphics and fonts	PostScript, TrueType
Sound	WAV, AVI, MP3, MIDI, WMA
Page description	PDF (Adobe Portable Document Format), HTML, XML
Video	Quicktime, MPEG-2, RealVideo, WMV

4

4

Internal Data Representation

- Reflects the
  - Complexity of input source
  - Type of processing required
- Trade-offs
  - Accuracy and resolution
    - Simple photo vs. painting in an art book
  - Compactness (storage and transmission)
    - More data required for improved accuracy and resolution
    - Compression represents data in a more compact form
    - Metadata: data that describes or interprets the meaning of data
- Ease of manipulation:
  - Processing simple audio vs. high-fidelity sound
- Standardization
  - Proprietary formats for storing and processing data (WordPerfect vs. Word)
  - De facto standards: proprietary standards based on general user acceptance (PostScript)

5

5

Data Types

- Numeric:
  - Used for mathematical manipulation
  - Add, subtract, multiply, divide
  - Types
    - Integer (whole number)
    - Real (contains a decimal point)
- Alphanumeric:
  - Characters: b T
  - Number digits: 7 9
  - Punctuation marks: ! ;
  - Special-purpose characters: $ &
- Numeric characters vs. numbers
  - Both entered as ordinary characters
  - Computer converts into numbers for calculation
    - Examples: Variables declared as numbers by the programmer (Salary$ in BASIC)
  - Treated as characters if processed as text
    - Examples: Phone numbers, ZIP codes

6

6

Alphanumeric Codes

- Arbitrary choice of bits to represent characters
  - Consistency:
    - input and output device must recognize same code
- Value of binary number representing character corresponds to placement in the alphabet
  - Facilitates sorting and searching
- Representing Characters
  - ASCII:
    - most widely used coding scheme
  - EBCDIC:
    - IBM mainframe (legacy)
  - Unicode:
    - developed for worldwide use

7

7

ASCII

- Developed by ANSI (American National Standards Institute)
- Represents
  - Latin alphabet, Arabic numerals, standard punctuation characters
  - Plus small set of accents and other European special characters
- ASCII
  - 7-bit code: 128 characters

8

8

ASCII Reference Table

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

74_{16} → 111 0100

9

9

EBCDIC

- Extended Binary Coded Decimal Interchange Code developed by IBM
  - Restricted mainly to IBM or IBM compatible mainframes
  - Conversion software to/from ASCII available

	ASCII	EBCDIC
Space	20_{16}	40_{16}
A	41_{16}	$C1_{16}$
b	62_{16}	82_{16}

  - Common in archival data
  - Character codes differ from ASCII

10

10

Unicode

- Most common 16-bit form represents 65,536 characters
- ASCII Latin-I subset of Unicode
  - Values 0 to 255 in Unicode table
- Multilingual: defines codes for
  - Nearly every character-based alphabet
  - Large set of ideographs for Chinese, Japanese and Korean
  - Composite characters for vowels and syllabic clusters required by some languages
- Allows software modifications for local-languages

11

11

Unicode Assignment Table

Code range (in hexadecimal)	
0000–007F	0000–00FF Latin-I (ASCII)
0080–0FFF	General character alphabets: Latin, Cyrillic, Greek, Hebrew, Arabic, Thai, etc.
1000–2FFF	Symbols and dingbats: punctuation, math, technical, geometric shapes, etc.
3000–3FFF	3000–33FF Miscellaneous punctuations, symbols, and phonetics for Chinese, Japanese, and Korean
4000–4DFF	Unassigned
4E00–9FFF	4E00–9FFF Chinese, Japanese, Korean ideographs
A000–AFFF	Unassigned
B000–BFFF	
C000–CFFF	AC00–D7AF Korean Hangui syllables
D000–DFFF	
E000–EFFF	Space for surrogates
F000–FBFF	E000–F8FF Private use
FC00–FFFF	FC00–FFFF Various special characters

12

12

2 Classes of Codes

- Printing characters
  - Produced on the screen or printer
- Control characters
  - Control position of output on screen or printer
    - VT: vertical tab LF: Line feed
  - Cause action to occur
    - BEL: bell rings DEL: delete current character
  - Communicate status between computer and I/O device
    - ESC: provides extensions by changing the meaning of a specified number of contiguous following characters

13

13

Control Code Definitions

NUL	(Null) No character; used to fill space	DLE	(Data Link Escape) Similar to escape, but used to change meaning of data control characters; used to permit sending of data characters with any bit combination
SOH	(Start of Heading) Indicates start of a header used during transmission	DC1,DC2, DC3, DC4	(Device Controls) Used for the control of devices or special terminal features
STX	(Start of Text) Indicates start of text during transmission	NAK	(Negative Acknowledgment) Opposite of ACK
ETX	(End of Text) Similar to above	SYN	(Synchronous) Used to synchronize a synchronous transmission system
EOT	(End of Transmission)	STB	(End of Transmission Block) Indicates end of a block of transmitted data
ENQ	(Enquiry) A request for response from a remote station; the response is usually an identification	CAN	(Cancel) Cancel previous data
ACK	(Acknowledge) A character sent by a receiving device as an affirmative response to a query by a sender	EM	(End of Medium) Indicates the physical end of a medium such as tape
BEL	(Bell) Rings a bell	SUB	(Substitute) Substitute a character for one sent in error
BS	(Backspace)	ESC	(Escape) Provides extensions to the code by changing the meaning of a specified number of contiguous following characters
HT	(Horizontal Tab)	FS, GS, RS, US	(File, group, record, and united separators) Used in optional way by systems to provide separations within a data set
LF	(Line Feed)	DEL	(Delete) Delete current character
VT	(Vertical Tab)		
FF	(Form Feed) Moves cursor to the starting position of the next page, form, or screen		
CR	(Carriage return)		
SO	(Shift Out) Shift to an alternative character set until SI is encountered		
SI	(Shift In) see above		

14

14

Keyboard Input

- Scan code
  - Two different scan codes on keyboard
    - One generated when key is struck and another when key is released
  - Converted to Unicode, ASCII or EBCDIC by software in terminal or PC
- Advantage
  - Easily adapted to different languages or keyboard layout
  - Separate scan codes for key press/release for multiple key combinations
    - Examples: shift and control keys

15

15

Other Alphanumeric Input

- OCR (optical character reader)
  - Scans text and inputs it as character data
  - Used to read specially encoded characters
    - Example: magnetically printed check numbers
  - General use limited by high error rate
- Bar Code Readers
  - Used in applications that require fast, accurate and repetitive input with minimal employee training
    - Examples: supermarket checkout counters and inventory control
  - Alphanumeric data in bar code read optically using wand
- Magnetic stripe reader:
  - alphanumeric data from credit cards
- Voice
  - Digitized audio recording common but conversion to alphanumeric data difficult
    - Requires knowledge of sound patterns in a language (phonemes) plus rules for pronunciation, grammar, and syntax

16

16

Image Data

- Photographs, figures, icons, drawings, charts and graphs
- Two approaches:
  - Bitmap or raster images of photos and paintings with continuous variation
  - Object or vector images composed of graphical objects like lines and curves defined geometrically
- Differences include:
  - Quality of the image
  - Storage space required
  - Time to transmit
  - Ease of modification

17

17

Bitmap Images

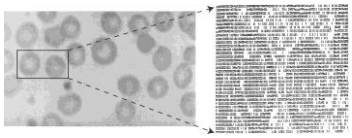
- Used for realistic images with continuous variations in shading, color, shape and texture
  - Examples:
    - Scanned photos
    - Clip art generated by a paint program
- Preferred when image contains large amount of detail and processing requirements are fairly simple
- Input devices:
  - Scanners
  - Digital cameras and video capture devices
  - Graphical input devices like mice and pens
- Managed by photo editing software or paint software
  - Editing tools to make tedious bit by bit process easier

18

18

Bitmap Images

- Each individual pixel (pi(x)cture element) in a graphic stored as a binary number
  - Pixel:
    - A small area with associated coordinate location
  - Example:
    - each point below represented by a 4-bit code corresponding to 1 of 16 shades of gray



19

19

Bitmap Display

- Monochrome:
  - black or white
    - 1 bit per pixel
- Gray scale:
  - black, white or 254 shades of gray
    - 1 byte per pixel
- Color graphics:
  - 16 colors, 256 colors, or 24-bit true color (16.7 million colors)
    - 4, 8, and 24 bits respectively

20

20

Storing Bitmap Images

- Frequently large files
  - Example: 600 rows of 800 pixels with 1 byte for each of 3 colors ➔ ~1.5MB file
- File size affected by
  - Resolution (the number of pixels per inch)
    - Amount of detail affecting clarity and sharpness of an image
  - Levels: number of bits for displaying shades of gray or multiple colors
    - Palette: color translation table that uses a code for each pixel rather than actual color value
- Data compression

21

21

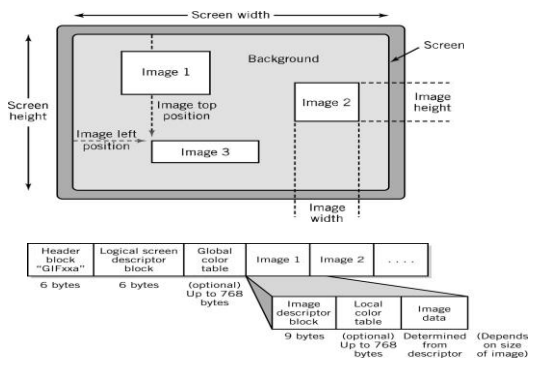
GIF (Graphics Interchange Format)

- First developed by CompuServe in 1987
- GIF89a enabled animated images
  - allows images to be displayed sequentially at fixed time sequences
- Color limitation: 256
- Image compressed by LZW (Lempel-Zif-Welch) algorithm
- Preferred for line drawings, clip art and pictures with large blocks of solid color
- Lossless compression

22

22

GIF (Graphics Interchange Format)



23

23

JPEG (Joint Photographers Expert Group)

- Allows more than 16 million colors
- Suitable for highly detailed photographs and paintings
- Employs lossy compression algorithm that
  - Discards data to decreases file size and transmission speed
  - May reduce image resolution, tends to distort sharp lines

24

24

Other Bitmap Formats

- TIFF (Tagged Image File Format): .tif (pronounced tif)
  - Used in high-quality image processing, particularly in publishing
- BMP (BitMaPped): .bmp (pronounced dot bmp)
  - Device-independent format for Microsoft Windows environment:
    - pixel colors stored independent of output device
- PCX: .pcx (pronounced dot p c x)
  - Windows Paintbrush software
- PNG: (Portable Network Graphics): .png (pronounced ping)
  - Designed to replace GIF and JPEG for Internet applications
  - Patent-free
  - Improved lossless compression
  - No animation support

25

25

Object Images

- Created by drawing packages or output from spreadsheet data graphs
- Composed of lines and shapes in various colors
- Computer translates geometric formulas to create the graphic
- Storage space depends on image complexity
  - number of instructions to create lines, shapes, fill patterns
- Movies Shrek and Toy Story use object images

26

26

Object Images

- Based on mathematical formulas
  - Easy to move, scale and rotate without losing shape and identity as bitmap images may
- Require less storage space than bitmap images
- Cannot represent photos or paintings
- Cannot be displayed or printed directly
  - Must be converted to bitmap since output devices except plotters are bitmap

27

27

Popular Object Graphics Software

- Most object image formats are proprietary
  - Files extensions include .wmf, .dxf, .mgx, and .cgm
- Macromedia Flash: low-bandwidth animation
- Micrographx Designer: technical drawings to illustrate products
- CorelDraw: vector illustration, layout, bitmap creation, image-editing, painting and animation software
- Autodesk AutoCAD: for architects, engineers, drafters, and design-related professionals
- W3C SVG (Scalable Vector Graphics) based on XML Web description language
  - Not proprietary

28

28

PostScript

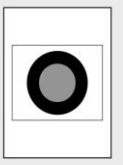
- Page description language:
  - list of procedures and statements that describe each of the objects to be printed on a page
    - Stored in ASCII or Unicode text file

```
288 396 translate        % move origin to center of page
0 0 144 0 360 arc        % define 2" radius black circle
fill

0.5 setgray              % define 1" radius gray circle
0 0 72 0 360 arc
fill

0 setgray                % reset color to black
-216 -180 moveto         % start at lower left corner
0 360 rmoveto            % and define rectangle
432 0 rmoveto            % ...one line at a time
0 -360 rmoveto
closepath                % completes rectangle
stroke                   % draw outline instead of fill

showpage                 % produce the image
```



- Interpreter program in computer or output device reads PostScript to generate image
- Scalable font support
  - Font outline objects specified like other objects

29

29

Representing Characters

- Characters stored in format like Unicode or ASCII
  - Text processed and stored primarily for content
- Presentation requirements like font stored with the character
  - Text appearance is primary factor
    - Example: screen fonts in Windows
- Glyphs:
  - Macintosh coding scheme that includes both identification and presentation requirement for characters

30

30

Bitmap vs. Object Images

Bitmap (Raster)	Object (Vector)
Pixel map	Geometrically defined shapes
Photographic quality	Complex drawings
Paint software	Drawing software
Larger storage requirements	Higher computational requirements
Enlarging images produces jagged edges	Objects scale smoothly
Resolution of output limited by resolution of image	Resolution of output limited by output device

31

31

Video Images

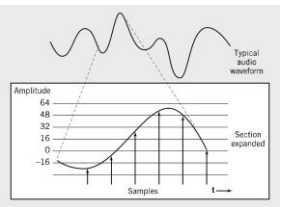
- Require massive amount of data
  - Video camera producing full screen 640 x 480 pixel true color image at 30 frames/sec ➔ 27.65 MB of data/sec
  - 1-minute film clip ➔ 1.6 GB storage
- Options for reducing file size: decrease size of image, limit number of colors, reduce frame rate
- Method depends on how video delivered to users
  - Streaming video: video displayed as it is downloaded from the Web server
    - Example: video conferencing
  - Local data (file on DVD or downloaded onto system) for higher quality
    - MPEG-2: movie quality images with high compression require substantial processing capability

32

32

Audio Data

- Transmission and processing requirements less demanding than those for video
- Waveform audio: digital representation of sound
  - Audio CD sampling rate = 44.1KHz



- Height of each sample saved as:
  - 8-bit number for radio-quality recordings
  - 16-bit number for high-fidelity recordings
    - 2 x 16-bits for stereo

33

33

MIDI

- MIDI (Musical Instrument Digital Interface):
  - instructions to recreate or synthesize sounds
    - Analog sound converted to digital values by A-to-D converter
  - Music notation system that allows computers to communicate with music synthesizers
  - Instructions that MIDI instruments and MIDI sound cards use to recreate or synthesize sounds.
    - Do not store or recreate speaking or singing voices
    - More compact than waveform
      - 3 minutes = 10 KB

34

34

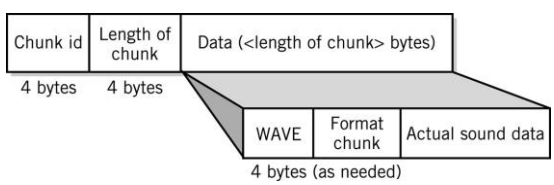
Audio Formats

- MP3
  - Derivative of MPEG-2 (ISO Moving Picture Experts Group)
  - Uses psychoacoustic compression techniques to reduce storage requirements
  - Discards sounds outside human hearing range: lossy compression
- WAV
  - Developed by Microsoft as part of its multimedia specification
  - General-purpose format for storing and reproducing small snippets of sound

35

35

.WAV Sound Format



36

36

Data Compression

- Compression: recoding data so that it requires fewer bytes of storage space.
- Compression ratio: the amount file is shrunk
- Lossless: inverse algorithm restores data to exact original form
  - Examples: GIF, PCX, TIFF
- Lossy: trades off data degradation for file size and download speed
  - Much higher compression ratios, often 10 to 1
    - Example: JPEG
  - Common in multimedia
- MPEG-2: uses both forms for ratios of 100:1

37

37

Compression Algorithms

- Repetition
  - **0** 5 8 7 **0 0 0 0** 3 4 **0 0 0** ➔ **0 1** 5 8 7 0 4 3 4 0 3
  - Example: large blocks of the same color
- Pattern Substitution
  - Scans data for patterns
  - Substitutes new pattern, makes dictionary entry

⌘	Pe	❋	pi	❖	ed
✹	er	●	ck	⌖	pe
✪	Pi				

- Example: 45 to 30 bytes plus dictionary
  - Peter Piper picked a peck of pickled peppers.
  - ⌘ t ✹ ✪ p ✹ ❋◎ ❖ a ⌖ ◎ of ❋ ◎l ❖ ⌖ **pp** ✹ s.

38

38

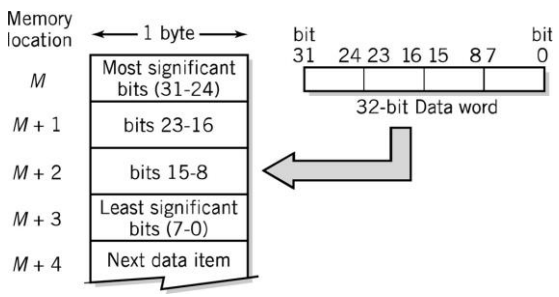
Internal Computer Data Format

- All data stored as binary numbers
  - Interpreted based on
    - Operations computer can perform
    - Data types supported by programming language used to create application
- Simple Data Types
  - Boolean:
    - 2-valued variables or constants with values of true or false
  - Char:
    - Variable or constant that holds alphanumeric character
  - Enumerated:
    - User-defined data types with possible values listed in definition
      - Type DayOfWeek = Mon, Tues, Wed, Thurs, Fri, Sat, Sun
  - Integer:
    - positive or negative whole numbers
  - Real :
    - Numbers with a decimal point, whose magnitude (large or small) exceeds computer's capability to store as an integer

39

39

Representing Numbers - 32-bit Data Word



40

40

Unsigned Numbers: Integers

- Unsigned whole number or integer
- Direct binary equivalent of decimal integer

4 bits: 0 to 9 16 bits: 0 to 9,999

8 bits: 0 to 99 32 bits: 0 to 99,999,999

Decimal	Binary	BCD
68	= 0100 0100 = $2^6 + 2^2 = 64 + 4 = $ ***68***	= 0110 1000 = $2^2 + 2^1 = $ ***6*** $2^3 = $ ***8***
99 (largest 8-bit BCD)	= 0110 0011 = $2^6 + 2^5 + 2^1 + 2^0 =$ $= 64 + 32 + 2 + 1 =$ ***99***	= 1001 1001 = $2^3 + 2^0$ $2^3 + 2^0$ = ***9*** ***9***
255 (largest 8-bit binary)	= 1111 1111 = $2^8 - 1 =$ ***255***	= 0010 0101 0101 = 2^1 $2^2 + 2^0$ $2^2 + 2^0$ = ***2*** ***5*** ***5***

41

41

Value Range: Binary vs. BCD

- BCD range of values < conventional binary representation
  - Binary: 4 bits can hold 16 different values (0 to 15)
  - BCD: 4 bits can hold only 10 different values (0 to 9)

No. of Bits	BCD Range		Binary Range	
4	0-9	1 digit	0-15	1+ digit
8	0-99	2 digits	0-255	2+ digits
12	0-999	3 digits	0-4,095	3+ digits
16	0-9,999	4 digits	0-65,535	4+ digits
20	0-99,999	5 digits	0-1 million	6 digits
24	0-999,999	6 digits	0-16 million	7+ digits
32	0-99,999,999	8 digits	0-4 billion	9+ digits
64	0-(10^{16}-1)	16 digits	0-16 quintillion	19+ digits

- Binary representation generally preferred
  - Greater range of value for given number of bits
  - Calculations easier
- BCD often used in business applications to maintain decimal rounding and decimal precision

42

42

Signed-Integer Representation

- 2's Complement

$$-2^{n-1}a_{n-1} + \sum_{i=0}^{n-2} 2^i a_i$$

+3 = 00000011 +2 = 00000010
+1 = 00000001 +0 = 00000000
-1 = 11111111 -2 = 11111110 -3 = 11111101

Range	-2^{n-1} through $2^{n-1} - 1$
Number of Representations of Zero	One
Negation	Take the Boolean complement of each bit of the corresponding positive number, then add 1 to the resulting bit pattern viewed as an unsigned integer.
Expansion of Bit Length	Add additional bit positions to the left and fill in with the value of the original sign bit.
Overflow Rule	If two numbers with the same sign (both positive or both negative) are added, then overflow occurs if and only if the result has the opposite sign.
Subtraction Rule	To subtract *B* from *A*, take the twos complement of *B* and add it to *A*.

43

43

Floating Point Representation (IEEE-754 fp)

$$N = (-1)^s \times 1.\text{fraction} \times 2^{(\text{biased exp.} - 127)}$$

- Sign: 1 bit
- Mantissa: 23 bits
  - We "normalize" the mantissa by dropping the leading 1 and recording only its fractional part
- Exponent: 8 bits
  - In order to handle both +ve and -ve exponents, we add 127 to the actual exponent to create a "biased exponent":
    - 2^{-127} => biased exponent = 0000 0000 (= 0)
    - 2^{0} => biased exponent = 0111 1111 (= 127)
    - 2^{+127} => biased exponent = 1111 1110 (= 254)

44

44