



CHAPTER 4: Data Formats

The Architecture of Computer Hardware, Systems Software & Networking: An Information Technology Approach

4th Edition, Irv Englander

John Wiley and Sons ©2010

PowerPoint slides authored by Wilson Wong, Bentley University

PowerPoint slides for the 3rd edition were co-authored with Lynne Senne, Bentley University

PowerPower slides modified by Gianluca Amato, Univ. di Chieti-Pescara



Data Formats

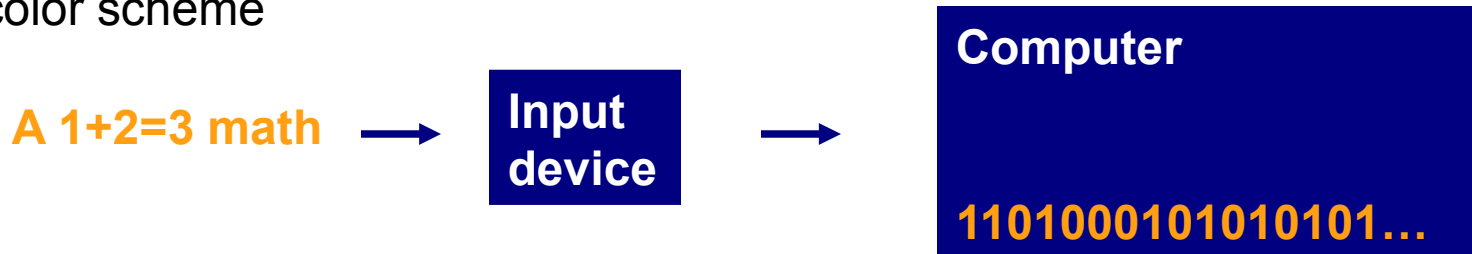
- Computers
 - Process and store all forms of data in binary format
- Human communication
 - Includes language, images and sounds
- Data formats:
 - Specifications for converting data into computer-usable form
 - Define the different ways human data may be represented, stored and processed by a computer



Sources of Data

- Binary/Digital 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

Figure 3.1 with this color scheme





Int. / ext. data representation

- Internal representation
 - Used internally by the program
 - Optimized for easy manipulation
- External representation
 - Used for storage and transmission
 - *Compression* represents data in a more compact form
 - *Metadata*: data that describes or interprets the meaning of data
 - Standardization
 - *Proprietary formats* for storing and processing data (WordPerfect vs. Word)
 - De facto standards: proprietary standards based on general user acceptance (PostScript)
 - Official standards



Some External Representations

Type of Data	Standard(s)
Alphanumeric	Unicode, ASCII, EDCDIC
Image (bitmapped)	<ul style="list-style-type: none">▪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



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)



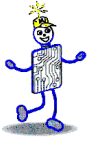
Data Types: Numeric

- Used for mathematical manipulation
 - Add, subtract, multiply, divide
- Types
 - Integer (whole number)
 - Real (contains a decimal point)
- Covered in Chapters 4 and 5



Data Types: Alphanumeric

- 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
 - Treated as characters if processed as text
 - Examples: Phone numbers, ZIP codes



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



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



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	ACJ	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

74₁₆

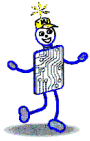
111 0100



EBCDIC

- Extended Binary Coded Decimal Interchange Code developed by IBM
 - Restricted mainly to IBM or IBM compatible mainframes
 - Conversion software to/from ASCII available
 - Common in archival data
 - Character codes differ from ASCII

	ASCII	EBCDIC
Space	20 ₁₆	40 ₁₆
A	41 ₁₆	C1 ₁₆
b	62 ₁₆	82 ₁₆



Collating Sequence

- Alphabetic sorting if software handles mixed upper- and lowercase codes
- In ASCII, numbers collate first; in EBCDIC, last
- ASCII collating sequence for string of characters

Letters

Adam A d a m

Adamian A d a m i a n

Adams A d a m s

Numeric Characters

1 011 0001

12 011 0001 011 0010

2 011 0010



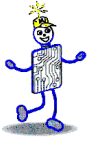
2 Classes of Codes

- *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



Unicode

- 32-bit code
- 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



Unicode

- Several representations
 - UCS-4
 - UTF-16
 - UTF-8



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



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
- *Bar Code Readers*
 - Used in applications that require fast, accurate and repetitive input with minimal employee training
 - Examples: supermarket checkout counters and inventory control
- *Magnetic stripe reader*: alphanumeric data from credit cards
- *RFID*: store and transmit data between RFID tags and computers
- *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



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



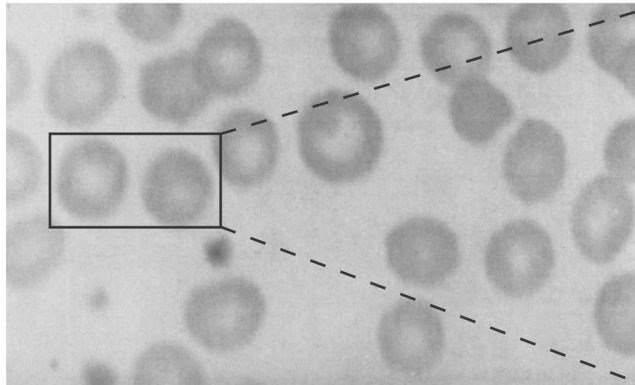
Bitmap Images

- 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



Bitmap Images

- Each individual *pixel* (*pic*ture *elem*ent) 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



```

24554531124743122111245788780788782222111111115558a654447a412 111 111 1 11111234568
2333322211474211 7456789986431 121 1 511 24578aaa9979aaa63211 11111121 12121474565
33334123567831 1 1211222435422 11 111111111122233445566778899aaa6443121222 11 2211112 4446483
222233 32777531 2 112111 12222331 1 111111111122212123444554442 1 1 1 122211 112338811
42225555744111111 1211 111111 12 211 2 111 1 121222222111 111 1211 1111211 1 244921221
744667888974112223212 11 11 11 1211 1212 1 1 21 11 111 11 1 12 11211 2 1 5094111234
498899999344211231111111 1111 111 2 121 122211 12 11 21 111 112 1222 1211 1 1284 122
aa9878654321211231344434332 1112331112323311 1 22121222111 21111111212211 4aa521221
657765432 1122212 1334455889a98752121 111211211 221 1 21121 1 1 2 2 1 1 11111 4a84 2345
9144113212112111344566666677874311 12 2221 1 111111 1 12 1 121212 11 111 2372125664
322221 1221 12113454431212 23456789111122111 1 122212111 1 3345444211 11 1111134442244333
312222222211211256421111121123678951 12211 211111 125789a987894311 1 11113841 33354
321113112 122248821 1121 11 124469421112332221 34589aaa9988998841 11138a41122333
32122232221 1257421 1111122219748531 11 1 1244a987322256678984311 11 38a4112445
31222222221124821 1222421 27982 122221 1157997831 121 1249a43121 1122a41 12445
21222222222222228841122243332112 11 24974 1221 2 1249a41111 212211 258752111222376711243
211213312113269731 12344333212111137853112112121 14949 1 21233122212112113a862212 249a4121259
21122221221 946781111211223222112 2485211 1 1 2 1278a51122445222211128a63211 1117a21244a
4123122222212768 1221221 21221122127452 1 1227a7121234222211122297421221 15974 121
55232 121133214a9113322224322111 59732 1 1 1594 2311342123212 13796a3 1 1379a952 1
22223 131 3222a922113421245322211 38942 111 12482 22344332222122214972 123112597a73 1
41222 13113127482 94312122312121 187511211117a72 2412431121212 111 5a671 112 2257a42
4123112212133a4 4 122222132112122 16a51 12211 17a71 24122 1 131212 4a431 11111 2389a4b
4133 1311222 4972 22222113112 12111289a2 1121 27a711312121 1221112121a94111211 1113597a6
413411311221 379731122122221121231 11399521 1321 16a72 1212121121122113a862212 1 12457b
3223112 1321248952112232222222 11468521 11111599413212 231 112211123a631221 111 12123
5123122112212369531 12332211111117a942 122222244123432112311 11 9a51 121 11 1
4 222122212212489a311113332212222 699a2 11211 1 359a2112321222321 117a9 11 122 1 111
412222221 1111238421 111122321116a82111111 1 247a951 1112222221116a821 111 1 28
5 1332222211112499a2 122211 2a494 121 1 13869521 111111 1578a4121111 222221
511221111222 14674511 111 123895 22111 1158a74111 11238a4aa5212221 3122 1 1
4222232 1 22211157884a334235789a82 11111221 11 1457995422232247aa952111 11 111121 11
412222222111121 15a789999999a8a3 111 1111 1335799877889a9a21 11 11 111111 11
41222222211 11121 1225a78999a9979421 11 12 1 1 111345689a9a4a7a7a2 1 1 1 111 121 11
32231 22122 1 11 235a777422222111211121 1 11156174a7878321 1 1 1111 221 1
97949117812 12112 1 11 222322211 121 1321 2211 112244224211111 11
  
```




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



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



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

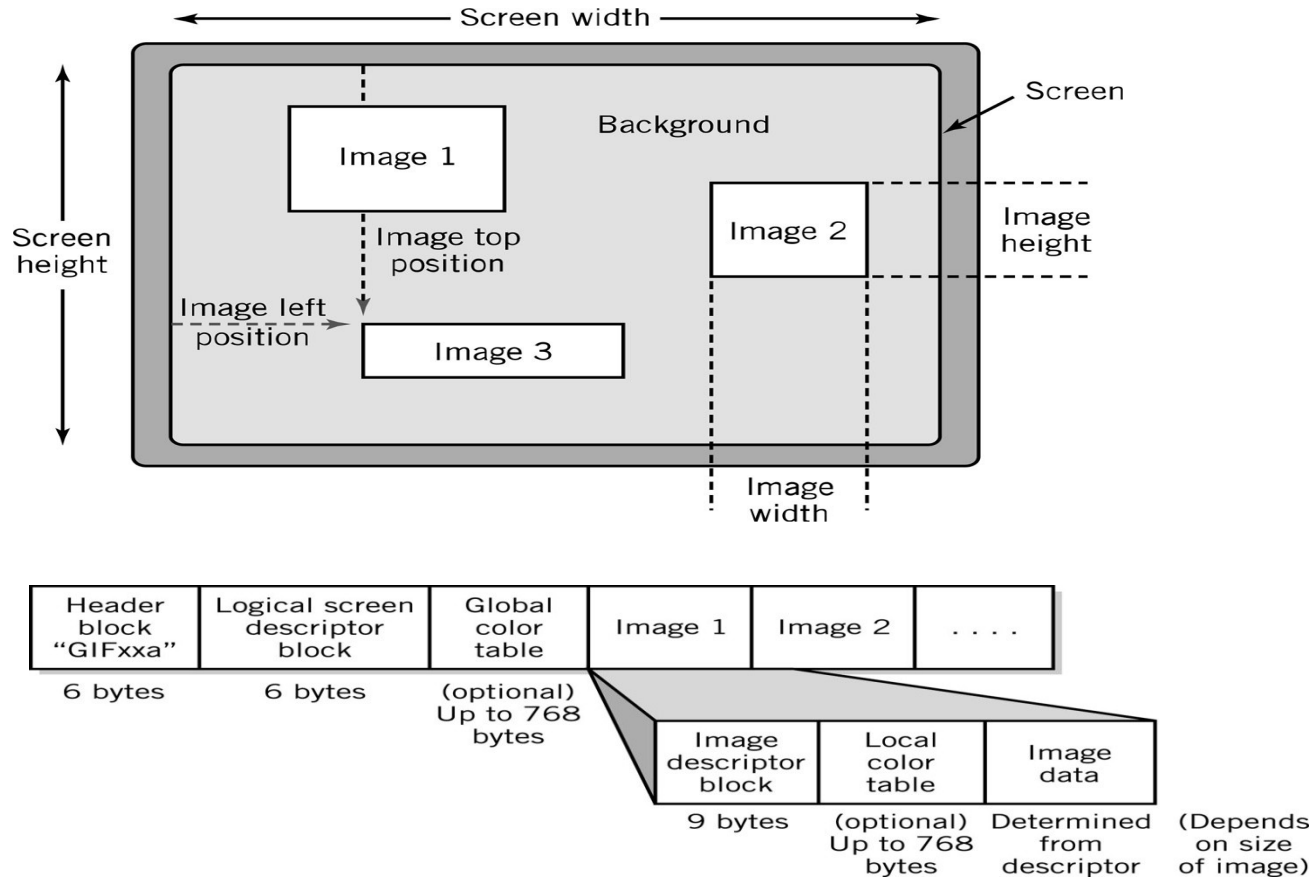


GIF (Graphics Interchange Format)

- 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*



GIF (Graphics Interchange Format)





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 decrease file size and transmission speed
 - May reduce image resolution, tends to distort sharp lines



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



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



SVG

- SVG (Scalable Vector Graphics) is a W3C standard for object images
 - XML based
 - Animations
 - Supported by web browsers



Bitmap vs. Object Images

Bitmap (Raster)

Pixel map

Photographic quality

Paint software

Larger storage requirements

Enlarging images produces jagged edges

Resolution of output limited by resolution of image

Object (Vector)

Geometrically defined shapes

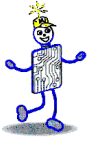
Complex drawings

Drawing software

Higher computational requirements

Objects scale smoothly

Resolution of output limited by output device



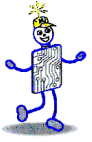
Video Images

- 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
 - Local data (file on DVD or downloaded onto system) for higher quality
 - MPEG-2: movie quality images with high compression require substantial processing capability

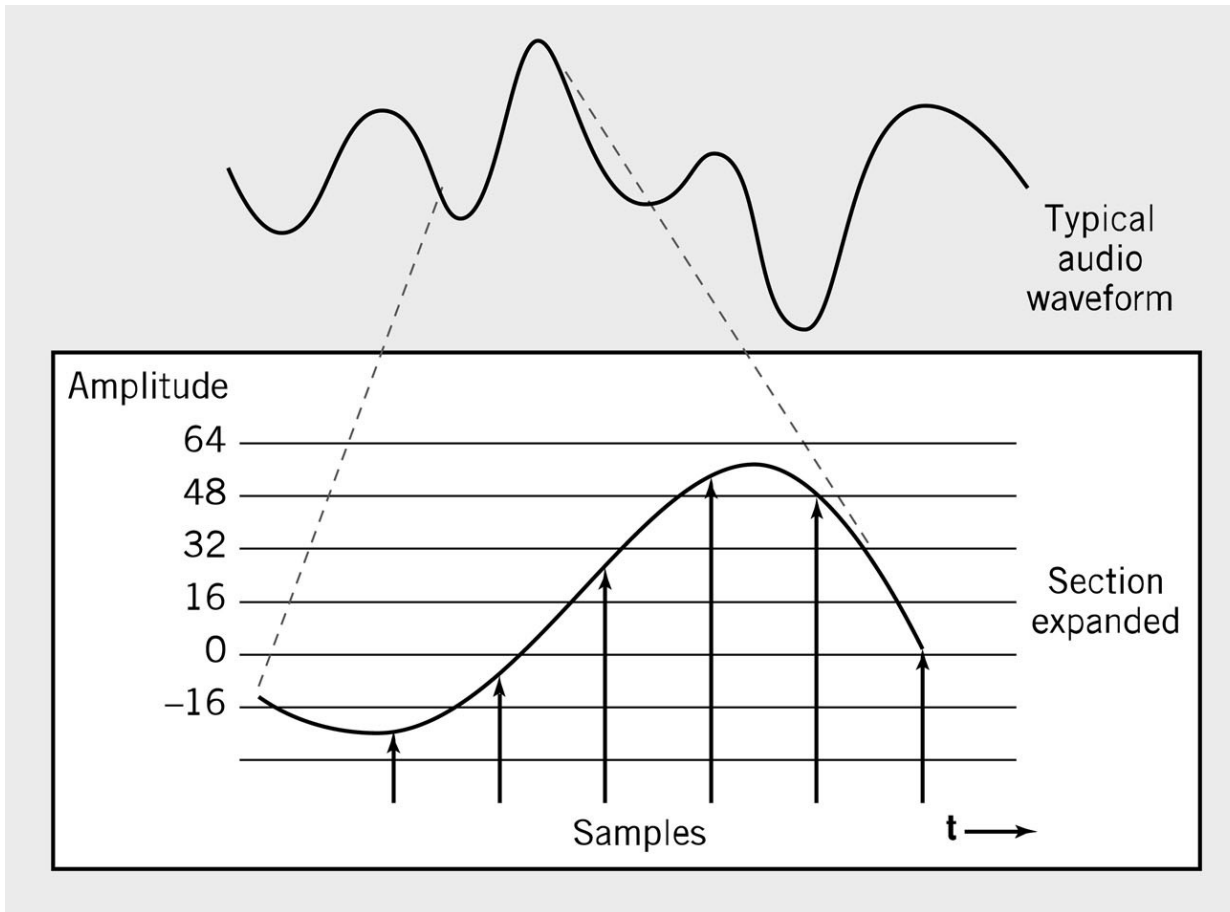


Audio Data

- Transmission and processing requirements less demanding than those for video
- *Waveform audio*: digital representation of sound
- *MIDI* (Musical Instrument Digital Interface): instructions to recreate or synthesize sounds
- Analog sound converted to digital values by *A-to-D converter*



Waveform Audio



Sampling rate normally 50KHz



Sampling Rate

- Number of times per second that sound is measured during the recording process.
 - 1000 samples per second = 1 KHz (kilohertz)
 - Example: 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

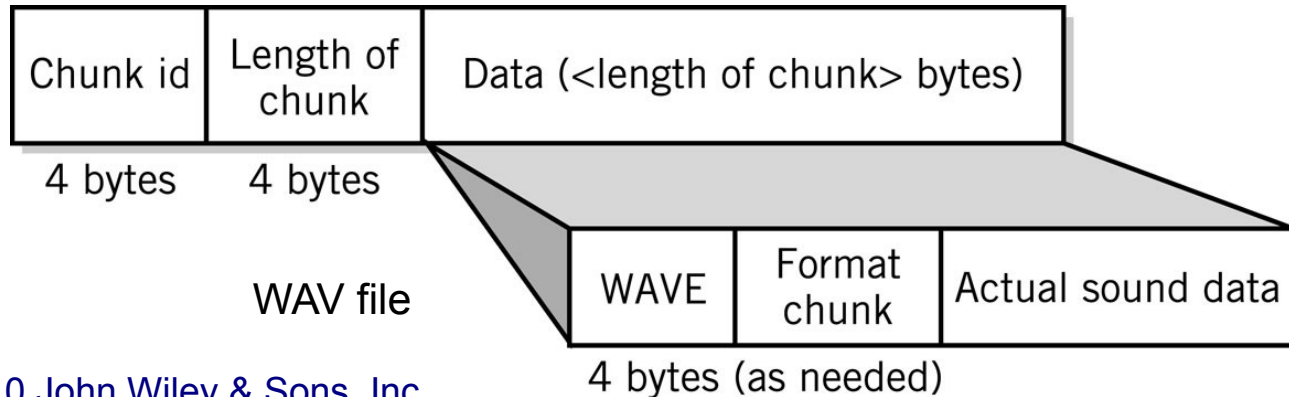
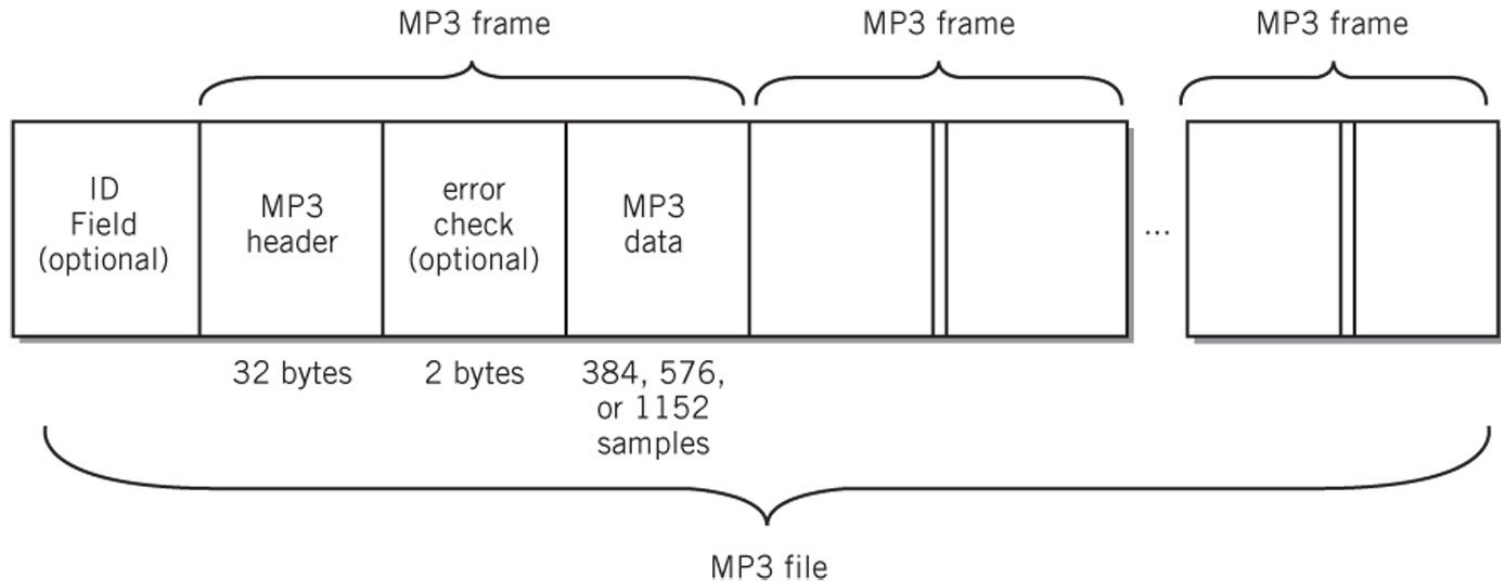


Audio Formats

- *MP3*
 - Derivative of MPEG-2 (ISO *M*oving *P*icture *E*xperts *G*roup)
 - Uses psychoacoustic compression techniques to reduce storage requirements
- *WAV*
 - Developed by Microsoft as part of its multimedia specification
 - General-purpose format for storing and reproducing small snippets of sound



Audio Data Formats





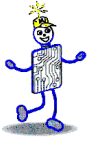
Page Description Languages

- Describe layout of objects on a displayed or printed page
- Objects may include text, object images, bitmap images, multimedia objects, and other data formats
- Examples
 - HTML, XHTML, XML
 - PDF
 - Postscript



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



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



5 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
 - Numbers whose magnitude, large or small, exceeds computer's capability to store as an integer



Copyright 2010 John Wiley & Sons

All rights reserved. Reproduction or translation of this work beyond that permitted in section 117 of the 1976 United States Copyright Act without express permission of the copyright owner is unlawful. Request for further information should be addressed to the Permissions Department, John Wiley & Sons, Inc. The purchaser may make back-up copies for his/her own use only and not for distribution or resale. The Publisher assumes no responsibility for errors, omissions, or damages caused by the use of these programs or from the use of the information contained herein.”