Click & Learn. Contents. WWW.MKDATA.DK

Now 205 pages of course material for self study or remote instruction. It would be to your advantage to print these pages. Click on the right hand frame, before you enter the print command. Welcome to the Click & Learn course. Used by schools teaching IT. Designed in several modules, some of which are sub divided, to facilitate reading. Please remember the **guest book.**

- **Please sign the guest book.**
- **See the guest book**
- **Or give your comments directly to: mkarbo@mkdata.dk**

Click & Learn. Module 1a. WWW.MKDATA.DK

About data

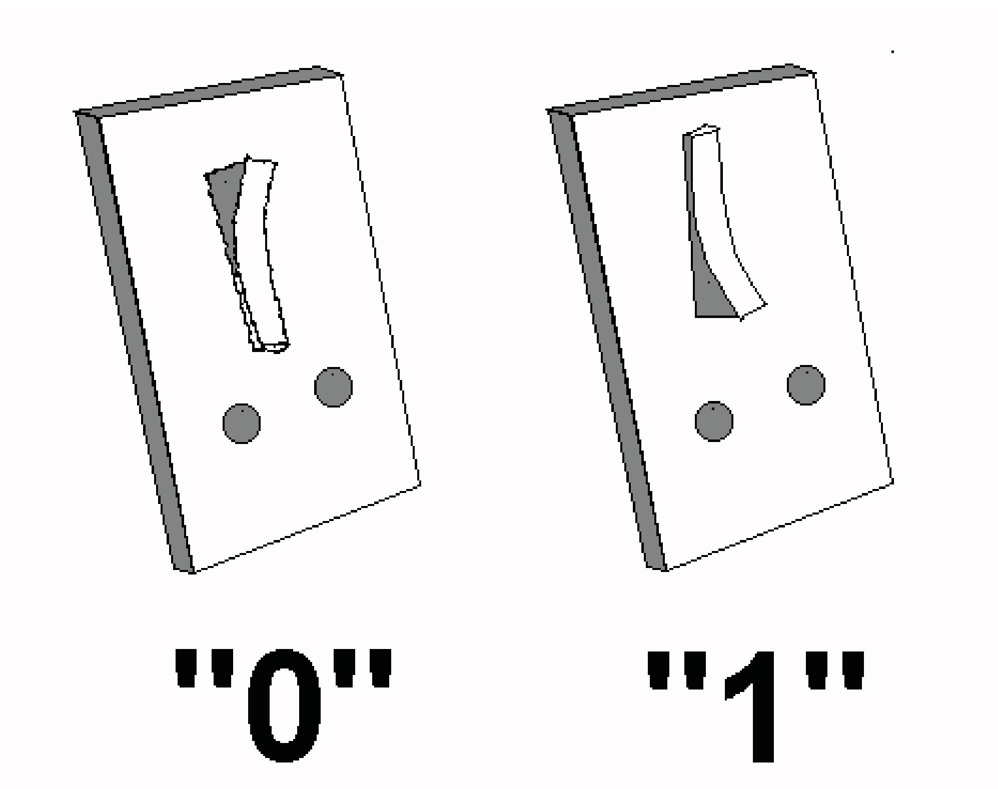
Our PC's are data processors. PC's function is simple: to process data, and the processing is done electronically inside the CPU and between the other components. That sounds simple, but *what* are data, and *how* are they processed electronically in a PC? That is the subject of these pages.

Analog data

The signals, which we send each other to communicate, are data. Our daily data have many forms: sound, letters, numbers, and other characters (handwritten or printed), photos, graphics, film. All these data are in their nature *analog*, which means that they are varied in their type. In this form, they are unusable in a PC. The PC can only process concise, simple data formats. Such data can be processed very effectively.

Digital data

The PC is an electric unit. Therefore, it can only deal with data, which are associated with electricity. That is accomplished using electric switches, which are either off or on. You can compare with regular household switches. If the switch if off, the PC reads numeral 0. If it is on, it is read as numeral one. See the illustration below:



With our electric switches, we can write 0 or 1. We can now start our data processing!

The PC is filled with these switches (in the form of transistors). There are literally millions of those in the electronic components. Each represents either a 0 or a 1, so we can process data with millions of 0's and 1's.

Bits

Each 0 or 1 is called a *bit*. Bit is an abbreviation of the expression **BI**nary digi**T**. It is called binary, since it is derived from the *binary number system:*

0	**1 bit**
1	**1 bit**
0110	**4 bit**
01101011	**8 bit**

The binary number system

The binary number system is made up of digits, just like our common *decimal* system (10 digit system). But, while the decimal system uses digits 0 through 9, the binary system only uses digits 0 and 1.

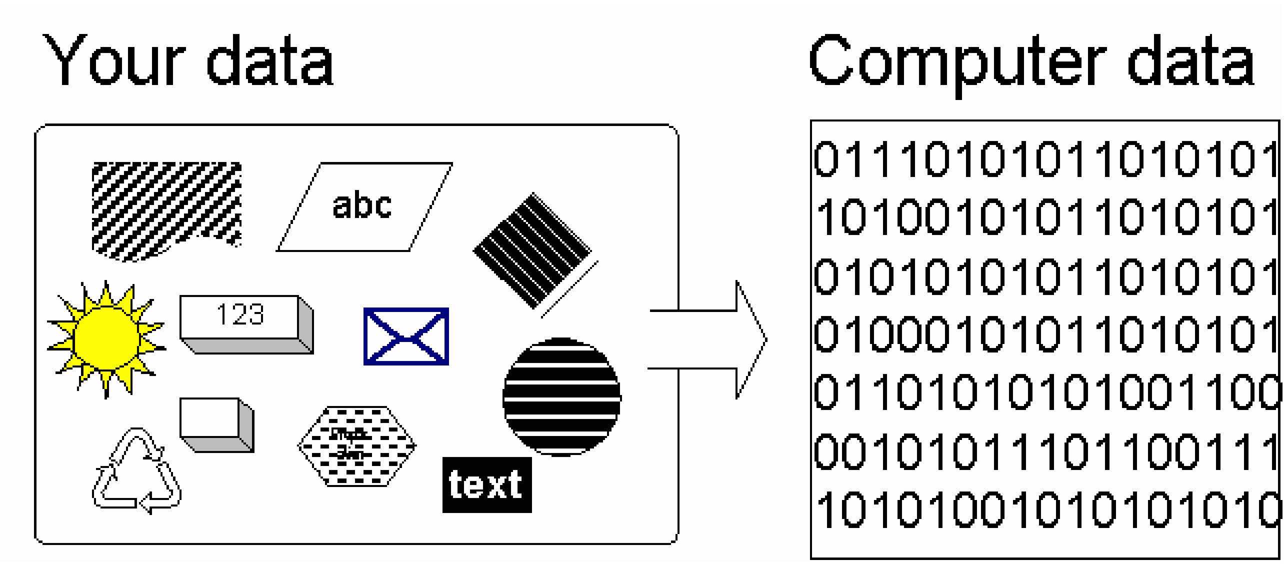
If you are interested in understanding the binary number system, then here is a brief course. Try if you can follow the system. See how numbers are constructed in the binary system, using only 0's and 1's:

Numbers, as known in the decimal-system	**Same numbers in binary system**
0	**0**
1	**1**
2	**10**
3	**11**
4	**100**
5	**101**
6	**110**
7	**111**
8	**1000**

Digital data

We have seen that the PC appears capable of handling data, if it can receive them as 0's and 1's. This data format is called digital. If we can translate our daily data from their analog format to digital format, they will appear as chains of 0's and 1's, then the PC can handle them.

So, we must be able to digitize our data. Pour text, sounds, and pictures into a funnel, from where they emerge as 0's and 1's:



Let us see how this can be accomplished**.**

Bytes

The most basic data processing is word processing. Let us use that as an example. When we do word processing, we work at a keyboard similar to a typewriter. There are 101 keys, where we find the entire alphabet A, B, C, etc. We also find the digits from 0 to 9 and all the other characters we need:,.-;():_?!"#*%&etc..

All these characters must be digitized. They must be expressed in 0's and 1's. Bits are organized in groups of 8. A group of 8 bits is called a byte.

8 bits = 1 byte, that is the system. Then, what can we do with bytes? First, let us see how many different bytes we can construct. A byte is an 8 digit number. We link 0's and 1's in a pattern. How many different ones can we make? Here is one: 01110101, and here is another: 10010101.

We can calculate that you can make 2 x 2 x 2 x 2 x 2 x 2 x 2 x 2 different patterns, since each of the 8 bits can have 2 values.

- **2^8** (two in the power of eight) is 256. Then there are 256 different bytes!

Now we assign a byte to each letter and other characters. And since we have 256 patterns to choose from, there is plenty of room for all. Here you see some examples of the "translation:"

Character	Bit pattern	Byte number		Character	Bit pattern	Byte number
A	01000001	65		¼	10111100	188
B	01000010	66		.	00101110	46

C	**01000011**	**67**		:	**00111010**	**58**
a	**01100001**	**97**		**$**	**00100100**	**36**
b	**01100010**	**98**		****	**01011100**	**92**
o	**01101111**	**111**		**~**	**01111110**	**126**
p	**01110000**	**112**		**1**	**00110001**	**49**
q	**01110001**	**113**		**2**	**00110010**	**50**
r	**01110010**	**114**		**9**	**00111001**	**57**
x	**01111000**	**120**		**©**	**10101001**	**169**
y	**01111001**	**121**		**>**	**00111110**	**62**
z	**01111010**	**122**		**‰**	**10001001**	**137**

When you write the word "summer", you write 6 letters. If the computer has to process that word, it will be digitized to 6 bytes. In other words, the word summer occupies 6 bytes in the PC RAM, when you type it, and 6 bytes on the hard disk, if you save it.

ASCII

ASCII means American Standard Code for Information Interchange. It is an industry standard, which assigns letters, numbers, and other characters within the 256 slots available in the 8 bit code.

The ASCII table is divided in 3 sections:

- Non printable system codes between 0 and 31.
- "Lower ASCII" between 32 and 127. This part of the table originates from older, American ADP systems, which work d on 7 bit character tables. Foreign letters, like Ø and Ü were not available then.
- "Higher ASCII" between 128 and 255. This part is programmable, in that you can exchange characters, based on which language you want to write in. Foreign letters are placed in this part.

Learn more about the ASCII table in Module 1b

An example

Let us imagine a stream of bits sent from the keyboard to the computer. When you type, streams of 8 bits are sent to the computer. Let
us look at a series of bits:

001100010011001000110011

Bits are combined into bytes (each 8 bits). These 24 bits are interpreted as three bytes. Let us read them as bytes: 00110001,
00110010, and 00110011.

When we convert these byte binary numbers to decimal numbers, you will see that they read as 49, 50, and 51 in decimal numbers. To
interpret these numbers, we have to look at the ASCII table. You will find that you have typed the numbers 1, 2, and 3.

About text and code

Now we have seen the PC's user data, which are always digitized. But there are many different kinds of data in the PC. You can differentiate between 2 fundamental types of data:

- **Program code**, which are data, that allow the PC to function.
- **User data,** like text, graphics, sound.

The fact is, that the CPU must have *instructions* to function. You can read more about this in the review of the CPU in module 3a. An instruction is a string of data, of 0's and 1's. The CPU is designed to recognize these instructions, which arrive together with the user input data to be processed.

The program code is thus a collection of instructions, which are executed one by one, when the program runs. Each time you click the mouse, or hit a key on the keyboard, instructions are sent from your software (program) to the CPU, telling it what to do next.

User data are those data, which tells the software how to respond. The letters, illustrations, home pages, etc., which you and I produce, are created with appropriate software.

Files

Both program code and user data are saved as *files on the* hard disk. Often, you can recognize the type of file by its suffix. Here are some examples:

Content	File name
Program code	START.EXE, WIN.COM, HELP.DLL, VMM32.VXD
User data	LETTER.DOC, HOUSE.BMP. INDEX.HTM

This is written as an introduction to naming files. The file name suffix determines how the PC will handle the file. You can read about this subject in some of my books, e.g. "DOS - teach yourself".

To overview.

Click & Learn. Module 1b. WWW.MKDATA.DK.

Character tables

Here you see the complete ASCII character table. First the part from ASCII-numbers 032 to 127:

ASCII-number	Common characters (in Windows)	Symbol	Wingdings
032			
033	!	!	!
034	"	∀	"
035	#	#	#
036	$	∃	$
037	%	%	%
038	&	&	&
039	'	∍	'
040	(	(	(
041	)	)	)
042	*	∗	*
043	+	+	+
044	,	,	,
045	-	−	-
046	.	.	.
047	/	/	/
048	0	0	0
049	1	1	1
050	2	2	2

051	3	3	3
052	4	4	4
053	5	5	5
054	6	6	6
055	7	7	7
056	8	8	8
057	9	9	9
058	:	:	:
059	;	;	;
060	<	<	<
061	=	=	=
062	>	>	>
063	?	?	?
064	@	≅	@
065	A	A	A
066	B	B	B
067	C	Χ	C
068	D	Δ	D
069	E	E	E
070	F	Φ	F
071	G	Γ	G
072	H	H	H
073	I	I	I
074	J	ϑ	J
075	K	K	K
076	L	Λ	L
077	M	M	M
078	N	N	N

079	**O**	Ο	O
080	**P**	Π	P
081	**Q**	Θ	Q
082	**R**	Ρ	R
083	**S**	Σ	S
084	**T**	Τ	T
085	**U**	Υ	U
086	**V**	ς	V
087	**W**	Ω	W
088	**X**	Ξ	X
089	**Y**	Ψ	Y
090	**Z**	Ζ	Z
091	**[**	[	[
092	****	∴	\\
093	**]**	]	]
094	**^**	⊥	^
095	**_**	_	_
096	**`**	‾	`
097	**a**	α	a
098	**b**	β	b
099	**c**	χ	c
100	**d**	δ	d
101	**e**	ε	e
102	**f**	φ	f
103	**g**	γ	g
104	**h**	η	h
105	**i**	ι	i
106	**j**	ϕ	j

107	k	κ	k
108	l	λ	l
109	m	μ	m
110	n	ν	n
111	o	ο	o
112	p	π	p
113	q	θ	q
114	r	ρ	r
115	s	σ	s
116	t	τ	t
117	u	υ	u
118	v	ϖ	v
119	w	ω	w
120	x	ξ	x
122	z	ζ	z
123	{	{	{
124	|	|	|
125	}	}	}
126	~	~	~
127			

Then the numbers from 0128 to 0255. Notice the leading zero.

I had problems with the width of the third column in the following table. Now it looks OK - thanks to Hans Rathje. HTML is tricky.

ASCII-number	Common characters (in Windows)	Symbol	Wingdings
0128	€		€

0129	**•**		•
0130	**‚**		‚
0131	***ƒ***		ƒ
0132	**„**		„
0133	**…**		…
0134	**†**		†
0135	**‡**		‡
0136	**ˆ**		ˆ
0137	**‰**		‰
0138	**Š**		Š
0139	**‹**		‹
0140	**Œ**		Œ
0141	**•**		•
0142	**Ž**		Ž
0143	**•**		•
0144	**•**		•
0145	**‘**		‘
0146	**’**		’
0147	**“**		“
0148	**”**		”
0149	**•**		•
0150	**–**		–
0151	**—**		—
0152	**˜**		˜
0153	**™**		™
0154	**š**		š
0155	**›**		›
0156	**œ**		œ

0157	•		•
0158	ž		ž
0159	Ÿ		Ÿ
0160		€	
0161	¡	ϒ	¡
0162	¢	′	¢
0163	£	≤	£
0164	¤	⁄	¤
0165	¥	∞	¥
0166	¦	ƒ	¦
0167	§	♣	§
0168	¨	♦	¨
0169	©	♥	©
0170	ª	♠	ª
0171	«	↔	«
0172	¬	←	¬
0173	-	↑	-
0174	®	→	®
0175	¯	↓	¯
0176	°	°	°
0177	±	±	±
0178	²	″	²
0179	³	≥	³
0180	´	×	´
0181	µ	∝	µ
0182	¶	∂	¶
0183	·	•	·
0184	¸	÷	¸

0185	**¹**	≠	¹
0186	**º**	≡	º
0187	**»**	≈	»
0188	**¼**	…	¼
0189	**½**	|	½
0190	**¾**	—	¾
0191	**¿**	↵	¿
0192	**À**	ℵ	À
0193	**Á**	ℑ	Á
0194	**Â**	ℜ	Â
0195	**Ã**	℘	Ã
0196	**Ä**	⊗	Ä
0197	**Å**	⊕	Å
0198	**Æ**	∅	Æ
0199	**Ç**	∩	Ç
0200	**È**	∪	È
0201	**É**	⊃	É
0202	**Ê**	⊇	Ê
0203	**Ë**	⊄	Ë
0204	**Ì**	⊂	Ì
0205	**Í**	⊆	Í
0206	**Î**	∈	Î
0207	**Ï**	∉	Ï
0208	**Ð**	∠	Ð
0209	**Ñ**	∇	Ñ
0210	**Ò**	®	Ò
0211	**Ó**	©	Ó
0212	**Ô**	™	Ô

0213	**Õ**	∏	Õ
0214	**Ö**	√	Ö
0215	**×**	⋅	×
0216	**Ø**	¬	Ø
0217	**Ù**	∧	Ù
0218	**Ú**	∨	Ú
0219	**Û**	⇔	Û
0220	**Ü**	⇐	Ü
0221	**Ý**	⇑	Ý
0222	**Þ**	⇒	Þ
0223	**ß**	⇓	ß
0224	**à**	◊	à
0225	**á**	〈	á
0226	**â**	®	â
0227	**ã**	©	ã
0228	**ä**	™	ä
0229	**å**	∑	å
0230	**æ**	⎛	æ
0231	**ç**	⎜	ç
0232	**è**	⎝	è
0233	**é**	⎡	é
0234	**ê**	⎢	ê
0235	**ë**	⎣	ë
0236	**ì**	⎧	ì
0237	**í**	⎨	í
0238	**î**	⎩	î
0239	**ï**	⎪	ï
0240	**ð**		ð

0241	**ñ**	〉	ñ
0242	**ò**	∫	ò
0243	**ó**	⌠	ó
0244	**ô**	⎮	ô
0245	**õ**	⌡	õ
0246	**ö**	⎞	ö
0247	**÷**	⎟	÷
0248	**ø**	⎠	ø
0249	**ù**	⎤	ù
0250	**ú**	⎥	ú
0251	**û**	⎦	û
0252	**ü**	⎫	ü
0253	**ý**	⎬	ý
0254	**þ**	⎭	þ
0255	**ÿ**		ÿ

To overview

Click & Learn. Module 2a. WWW.MKDATA.DK

Remember, you can print these pages. Press Ctrl+p, [Enter], then you get a hard copy. These pages go together with modules 2b, 2c, 2d and 2e. More than 30 pages about:

Introduction to the PC

The technical term for a PC is *micro data processor*. That name is no longer in common use. However, it places the PC in the bottom of the computer hierarchy:

- **Mainframes** are the very largest computers - million dollar machines, which can occupy more than one room, An example is IBM model 390.
- **Minicomputers** are large powerful machines. They typically serve a network of simple terminals. IBM's AS/400 is an example of a minicomputer.
- **Workstations** are powerful user machines. They have the power to handle complex engineering applications. They use the UNIX or sometimes the NT operating system. Workstations can be equipped with powerful RISC processors like Digital Alpha or MIPS.
- **PC's** are the Benjamin's in this order: Small inexpensive, mass produced computers. They work on DOS, Windows, or similar operating systems. They are used for standard applications.

The point of this history is, that Benjamin has grown. He has actually been promoted to captain! Today's PC's are just as powerful as minicomputers and mainframes were not too many years ago. A powerful PC can easily keep up with the expensive workstations. How have we advanced this far?

The PC's success

The PC came out in 1981. In less than 20 years, it has totally changed our means of communicating. When the PC was introduced by IBM, it was just one of many different micro data processors. However, the PC caught on. In 5-7 years, it conquered the market. From being an IBM compatible PC, it became *the standard.*

If we look at early PC's, they are characterized by a number of features. Those were instrumental in creating the PC success.

- The PC was from the start standardized and had an open *architecture.*
- It was well documented and had great possibilities for expansion.
- It was inexpensive, simple and robust (definitely not advanced).

The PC started as IBM's baby. It was their design, built over an Intel processor (8088) and fitted to Microsoft's simple operating system MS-DOS.

Since the design was well documented, other companies entered the market. They could freely copy the central system software (BIOS) and the ISA bus, since they were not patented. Slowly, a myriad of companies developed, manufacturing IBM compatible PC's and components for them.

The Clone was born. A clone is a copy-machine. A machine, which can do precisely the same as the original (read *Big Blue* - IBM). Some of the components (for example the hard disk) may be identical to the original. However, the Clone has another name (Compaq, Olivetti, etc.), or it has no name at all. This is the case with "the real clones." Today, we differentiate between:

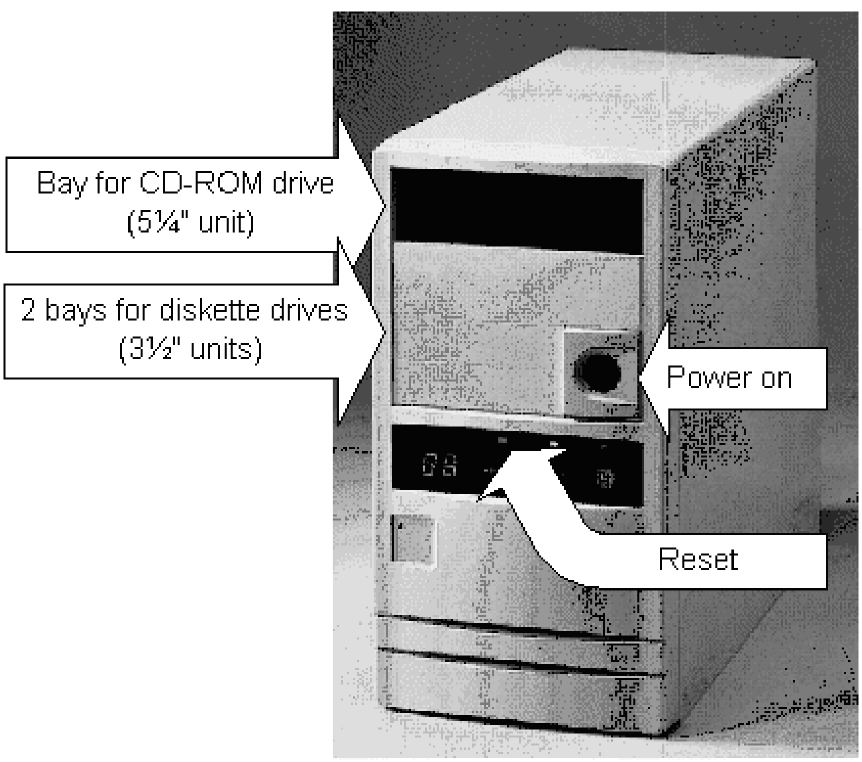
- Brand names, PC's from IBM, Compaq, AST, etc. Companies which are so big, so they develop their own hardware components.
- Clones, which are built from standard components. Anyone can make a clone.

Since the basic technology is shared by *all* PC's, I will start with a review of that.

The PC construction

The PC consists of a *central unit* (referred to as the computer) and various peripherals. The computer is a box, which contains most of the working electronics. It is connected with cables to the peripherals.

On these pages, I will show you the computer and its components. Here is a picture of the computer:



Here is a list of the PC components. Read it and ask yourself what the words mean.. Do you recognize all these components? They will be covered in the following pages.

Components in the central unit - the computer	Peripherals
The main board: CPU, RAM, cache, ROM chips with BIOS and start-up programs. Chip sets (controllers). Ports, busses and expansion slots. **Drives**: Hard disk(s), floppy drive(s), CD-ROM, etc. **Expansion cards**: Graphics card (video adapter), network controller, SCSI controller. Sound card, video and TV card. Internal modem and ISDN card	Keyboard and mouse Monitor Printer Scanner Tape drives, etc. External modem

So, how are the components connected. What are their functions, and how are they tied together to form a PC? That is the subject of Click and Learn. So, please continue reading...

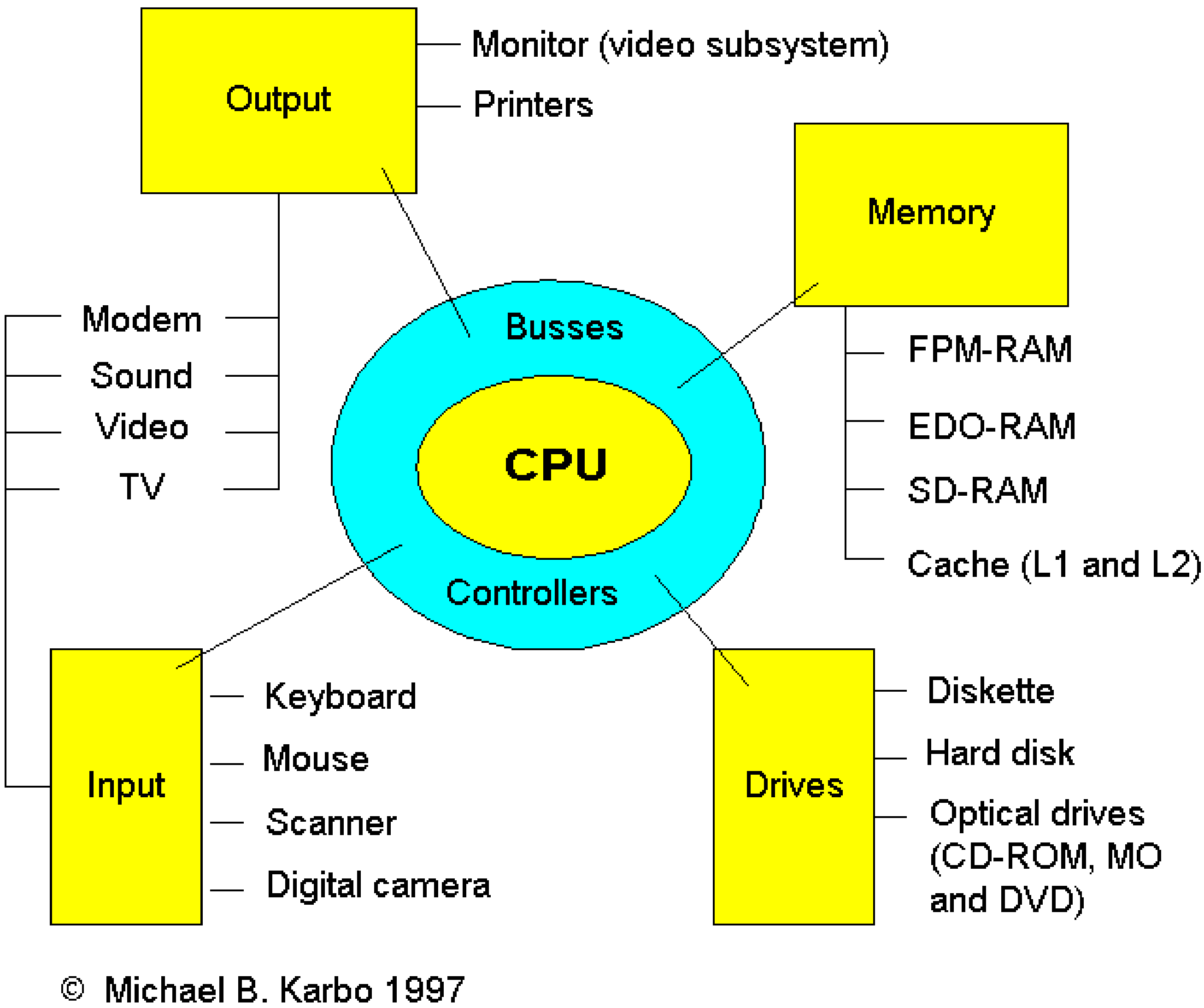
History of the PC

Computers have their roots 300 years back in history. Mathematicians and philosophers like Pascal, Leibnitz, Babbage and Boole made the foundation with their theoretical works. Only in the second half of this century was electronic science sufficiently developed, to make practical use of their theories.

The modern PC has roots back to USA in the 1940's. Among the many scientists, I like to remember John von Neumann (1903-57). He was a mathematician, born in Hungary. We can still use his computer design today. He broke computer hardware down in five primary parts:

- CPU
- Input
- Output
- Working memory
- Permanent memory

Actually, von Neumann was the first to design a computer with a working memory (what we today call RAM). If we apply his model to current PC's, it will look like this:



All these subjects will be covered.

Data exchange - the mainboard

It is a printed circuit board, on which multiple chips, ports (plug ins), and other electronic components are mounted. In the PC, data are exchanged continuously between these components. Therefore it is important to understand each component, its connections and characteristics. All data exchange is done on the *system board*, which thus is the most important component in the PC. So, now we will start with a more technical evaluation of the system board.

The mainboard components

The PC is built around the main, system or *mother* board (all meaning the same). This board is so essential for the PC, because it holds the CPU and all its connections. Let us see, what you can find on it:

- **ROM-chips with BIOS and other programs**
- **CMOS, storing system setup data**
- **The CPU**
- **L2-cache**
- **Chip sets with I/O controllers**

- **RAM (*Random Access Memory)* mounted in SIMM or DIMM chips**
- **Cards to connect with keyboard and mouse**
- **Serial and parallel ports**
- **Connectors to disk drives and EIDE drive (hard disk, CD-ROM etc.)**
- **Slots for expansion cards**
- **Jumpers to adjust voltage, system bus speed, clock, etc.**
- **Contacts to reset HD activity, speaker, etc.**

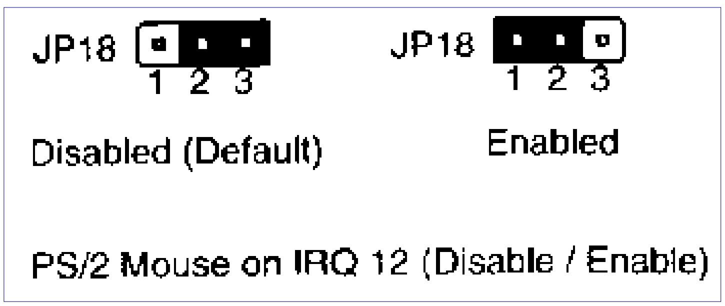
I want to describe many of these gismos and components on the following pages.

Use the manual

If you are interested in the system board and the technical aspects of the PC, the system board *manual* is an essential tool. With patience, you can find much information there. It is especially valuable, if you understand the system board principles (clock factor, bus speed, etc.).

For example, you can read how to set jumper switches to utilize some options.

I have connected a PS/2-mouse to the special AUX-port. It is a small connector on the system board, where I can connect a mini DIN connector. In that way, I have connected the mouse, without occupying any COM ports. In this situation, I have to reset a jumper switch. My manual tells me it is JP18:



The PC start-up process

When you turn power on, several things happen in the PC:

- You hear the fan motor starting. There are one or more cooling fans in the PC. They produce a whirring sound.
- After a few seconds, text starts to scroll on the screen.
- Now the PC tests and counts the RAM. You see a number on the screen. It increases in size.

To understand the working of the PC, it is useful to study the PC start-up process. Those are events, which take place from power-on until the PC is ready to work. Remember, the PC can do nothing without receiving instructions. These instructions are commands, which are sent to the CPU. During start-up, the PC reads the commands in this sequence:

- First it receives commands from the ROM chips. Those chips are inherent in any computer. They contain the POST and BIOS instructions, which we will look at shortly.
- Next, the operating system is read from the *hard disk* (or from floppy drive A). This is called the boot process.

The ROM chips

ROM (*Read Only Memory)*. The ROM chips are on the system board. They contain. *system software*. System software are instructions, which enable the PC to coordinate the functions of various computer components.

The ROM chips contain instructions, which are specific for that particular system board. Those instructions will remain in the PC throughout its life. They will usually not be altered. Primarily, they are *start-up instructions.* There are different parts in the start-up instructions. For most users, they are all woven together. You can differentiate between:

- **POST** (Power On Self Test)
- **The *Set-up* instructions**, which connect with the CMOS instructions
- **BIOS instructions**, which connect with the various hardware peripherals
- **The *Boot* instructions**, which calls the operating system (DOS, OS/2, or Windows)

All these instructions are in ROM chips, and they are activated on by one during start-up. Let us look at each part.

POST

Power On Self Test is the first instruction executed during start-up. It checks the PC components and that everything works. You can recognize it during the RAM test, which occurs as soon as you turn power on.

As users, we have only limited ability to manipulate the POST instructions. But certain system boards enable the user to order a quick system check. Some enable the user to disable the RAM test, thereby shortening the duration of the POST. The duration of the POST can vary considerably in different PC's. On the IBM PC 300 computer, it is very slow. But you can disrupt it by pressing [Esc].

If POST detects errors in the system, it will write error messages on the screen. If the monitor is not ready, or if the error is in the video card, it will also sound a pattern of beeps (for example 3 short and one long) to identify the error to the user. If you want to know more of the beeps, you can find explanations on the Award, AMI and Phoenix web sites.

POST also reads those user instructions, which are found in CMOS:

CMOS

CMOS (Complimentary Metal Oxide Semiconductor) is a small amount of memory in a special RAM chip. Its memory is maintained with electric power from a small battery. Certain system data are stored in this chip. They must be read to make the PC operable. There may be 100 to 200 bytes of data regarding date, time, floppy and hard disk drives, and much more.

CMOS data can be divided in two groups:

- **Data, which POST can not find during the system test.**
- **Data, which contain user options.**

For example, POST cannot by itself find sufficient information about the floppy drive(s). Floppy drives are so "dumb," that POST cannot read whether they are floppy drives or not, nor what type. About the same goes for IDE hard disks, while EIDE hard disks are a little more "intelligent," However, POST still needs assistance to identify them 100% correctly.

The same goes for RAM: POST can count how much RAM is in the PC. However, POST cannot detect whether it is FPM, EDO or SD RAM. Since the CPU and BIOS reads data from RAM chips differently, depending on the RAM type, that type must be identified.

Configuration

The PC must be configured, be supplied with this information. That is done in the factory or store, where it is assembled. This information is stored in CMOS, where they stay. CMOS data only need to be updated, when different or additional hardware components are installed. This could be a different type hard disk or floppy disks or an new RAM type, Often he user can do this.

Other data in CMOS contain various *user options*. Those are data, which you can write to CMOS. For example, you can adjust date and time, which the PC then adjusts every second. You can also choose between different system parameters. Maybe you want a short system check instead of a long one. Or if you want the PC to try to boot from hard disk C before trying floppy disk A, or vice versa. These options can be written to CMOS.

Many of the options are of no interest to the ordinary user. These are options, which regard controller chips on the system board, which *can* be configured in different ways. Ordinarily, there is no need to make such changes. The system board manufacturer has already selected the optimal configurations. They recommend in their manuals, that you do not change these *default* settings.

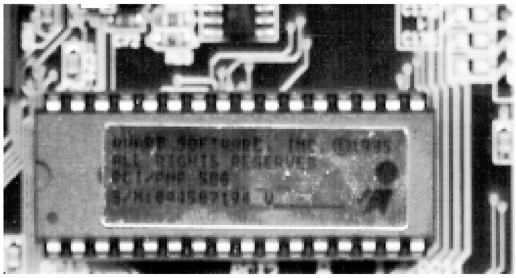
We can conclude, that CMOS data are essential system data, which are vital for operation of the PC. Their special feature is, that they are user adjustable. Adjustments to CMOS are made during start-up.

Suppliers of system software

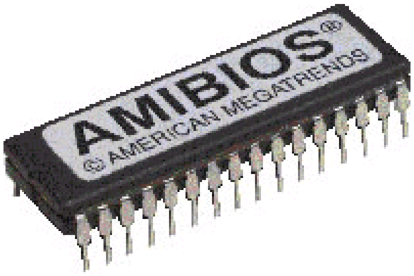
All PC's have instructions in ROM chips on the system board. The ROM chips are supplied by specialty software manufacturers, who make BIOS chips. The primary suppliers are:

- **Phoenix**
- **AMI (*American Megatrends*)**
- **Award**

You can read the name of your BIOS chip during start-up. You can also see the chip on the system board. Here is a picture (slightly blurred) of an Award ROM chip:



Here is an AMI chip with BIOS and start-up instructions:



The Setup program

You communicate with the BIOS programs and the CMOS memory through the so-called Setup program.

Typically you reach the Setup program by pressing [Delete] immediately after you power up the PC. That brings you to a choice of setup menus. You leave Setup by pressing [Esc], and choose "Y" to restart the PC with the new settings. Generally, you should not change these settings, unless you know precisely what you are doing.

The Setup program can do many things for you. You *have* to enter Setup, if you install a different type or additional disk drive in your PC. Certain BIOS's will also need adjustment of its settings, if a CD ROM drive is installed on one of the EIDE channels.

Modifying the boot sequence

You can change the boot sequence from A:, C: to C:, A:. That means, that the PC will not try to boot from any diskette in the A drive. That will protect you from certain virus attacks from the boot sector. Also, the boot process will not be blocked by any diskette in the A drive. If you need to boot from A-drive (for example, if you want to install Windows 97), you have to enter Set-up again, and change the boot sequence to A:, C:. That is no problem.

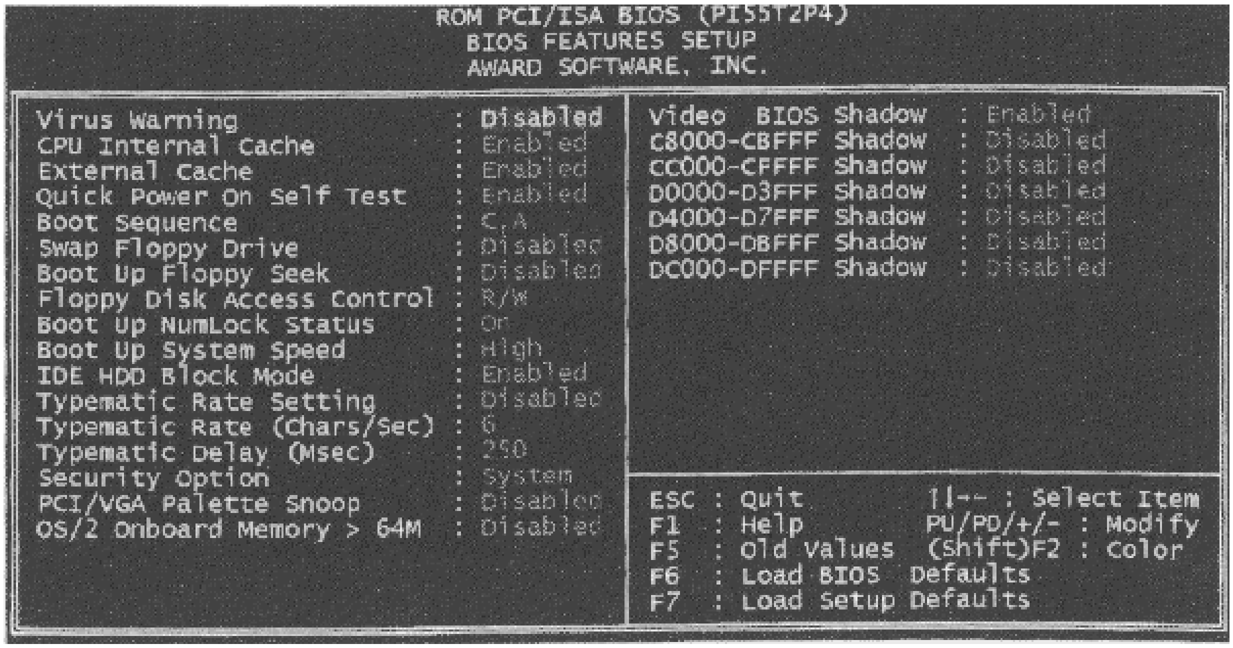
Power Management

You also use the Setup program to regulate *Power Management*, which is the power saving features in the system board. For example, you can make the CPU shut down after one minute of no activity. There are plenty of settings available in this area.

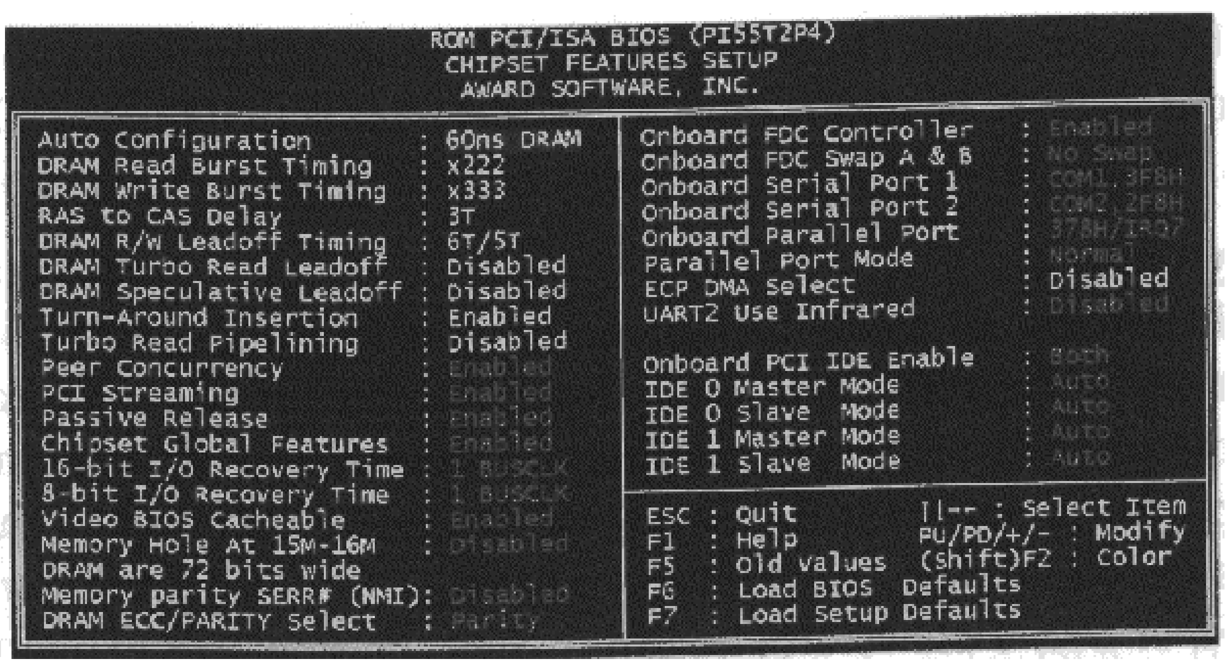
Password Protection

You protect the Setup program with a password. This is used widely in schools, where they do not want the little nerds to make changes in the setup. Please remember the password (write it down in the mainboard manual). If you forget it you have to remove the battery from the mainboard. Then all user-input to the CMOS is erased - including the password.

Here is a scanned image from a Setup program. It belongs to my favorite board (from ASUS). Here you see the "BIOS Feature Setup," where you can select start-up choices:



Here we are are in the special "Chip set Feature Setup." These choices relate to the chip sets and, most likely, need no changes:



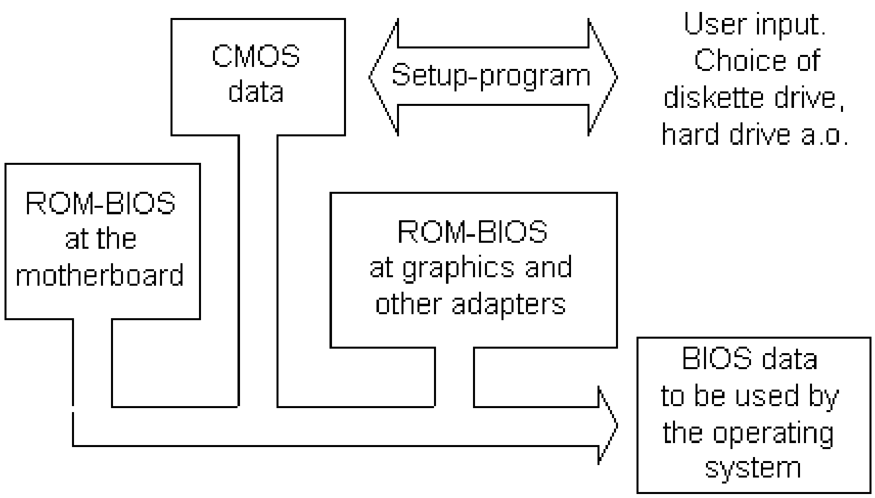
The BIOS programs

During start-up. the BIOS programs are read from a ROM chip. BIOS is abbreviation of *Basic Input Output System* and those are programs, which are linked to specific hardware systems. For example, there is a BIOS routine, which identifies how the PC reads input from the keyboard.

BIOS is a typical link in the IBM compatible PC design. The BIOS programs control hardware, the user (programmer) controls hardware via a call to BIOS.

BIOS typically occupy 1 MB, and the programs are saved ROM chips on the system board.

During start-up, BIOS is read from ROM chips. That information is supplemented with the system data saved in CMOS. Furthermore, there is BIOS code on the expansion cards. The expansion cards are external hardware, as interpreted by the system board, and the BIOS code, which is linked to the expansion card, must be included in the configuration. Therefore, this *expansion card ROM* is read during start-up, and the program code is woven together with other BIOS data. It is all written into RAM, where it is ready for the operating system, as you can see here:



Otherwise, the BIOS routines are not always in use. They can be regarded as basic program layers in the PC. Many programs routinely bypass BIOS. In that case, they "write direct to hardware", as we say. Windows contains program files, which can be written directly to all kinds of hardware - bypassing BIOS routines. One example is the COM ports. If you use the BIOS routines connected with them, you can transmit only at max. 9600 baud on the modem. That is insufficient. Therefore, Windows will assume control over the COM port.

BIOS-update

BIOS programs can be *updated*. The modern system board has the BIOS instructions in *flash-ROM,* which can be updated. You can get new BIOS-software from your supplier or on the Internet, which can be read onto the system board. The loading is a special process, where you might need to change a *jumper switch* on the system board. Usually, you do not need to do this, but it is a nice available option.

ATX

The latest PC electronic standard is called ATX. It consists of a new type system board with a specific physical design smaller than the traditional board (30.5 cm X 19 cm). The I/O connectors COM1, COM2 and LPT, keyboard, mouse and USB are mounted directly on the system board. The ATX board requires specifically designed chassis's with an I/O access opening measuring 1¾ by 6¼ inch. ATX is designed by Intel, but has gained general acceptance.

The ATX system board is more "intelligent" than the ordinary type. In a few years, it will be wide spread. It includes advanced control facilities, where the BIOS program continually checks the CPU temperature and voltages, the cooling fans RPM, etc. If over heating occurs, the PC will shut down automatically. The PC can also be turned on by for example modem signals, since the power supply is controlled by the system board. The on/off button will turn the PC "down" without turning it completely off.

If you want a PC designed for the future, the ATX layout is what you should go for.

Read more about the boot process and system bus in Module 2b

Read more about I/O busses in module 2c

Read more about the system board chip set in module 2d

Read more about RAM in module 2e

I can recommend two books for further studies. Gunnar Forst: "PC Principals", from MIT is excellent. Also "The Winn L. Rosch Hardware Bible" from Brady covers the same subjects.

To overview

Click & Learn. Module 2b. WWW.MKDATA.DK

About the System Bus

In this module, you can read about the following subjects, which add to our tour of the PC:

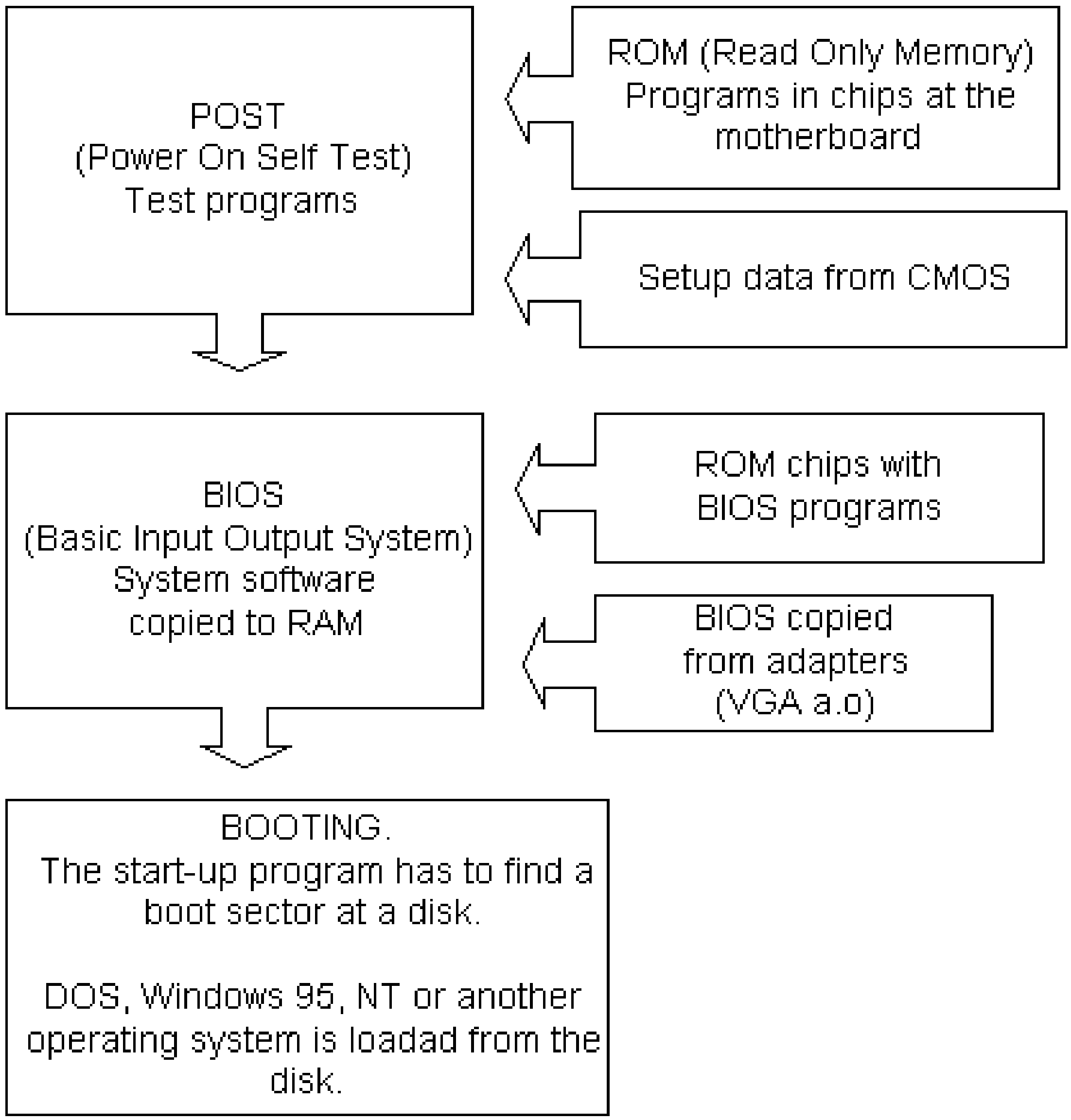
The boot process

The last step in the PC start-up is reading the operating system. The start-up program is instructed to find the *Master boot sector*. The boot sector is the very first sector on either hard disk (C) or floppy drive A.

By default, the PC will look for a boot sector in floppy drive A. That is why the PC "drops dead" if there is a different diskette in A drive. If there is no diskette in A drive, the start-up program will search for the boot sector on hard drive C. When the boot sector is found, a small program segment (*boot-strap*) is read from there. The boot-strap then takes over control of the PC. The start-up program has done its job. Now DOS, Windows, or another operating system takes control.

Read more about boot sectors, etc. in module 6a, which deals with file systems.

Here is an illustration of the start-up process:

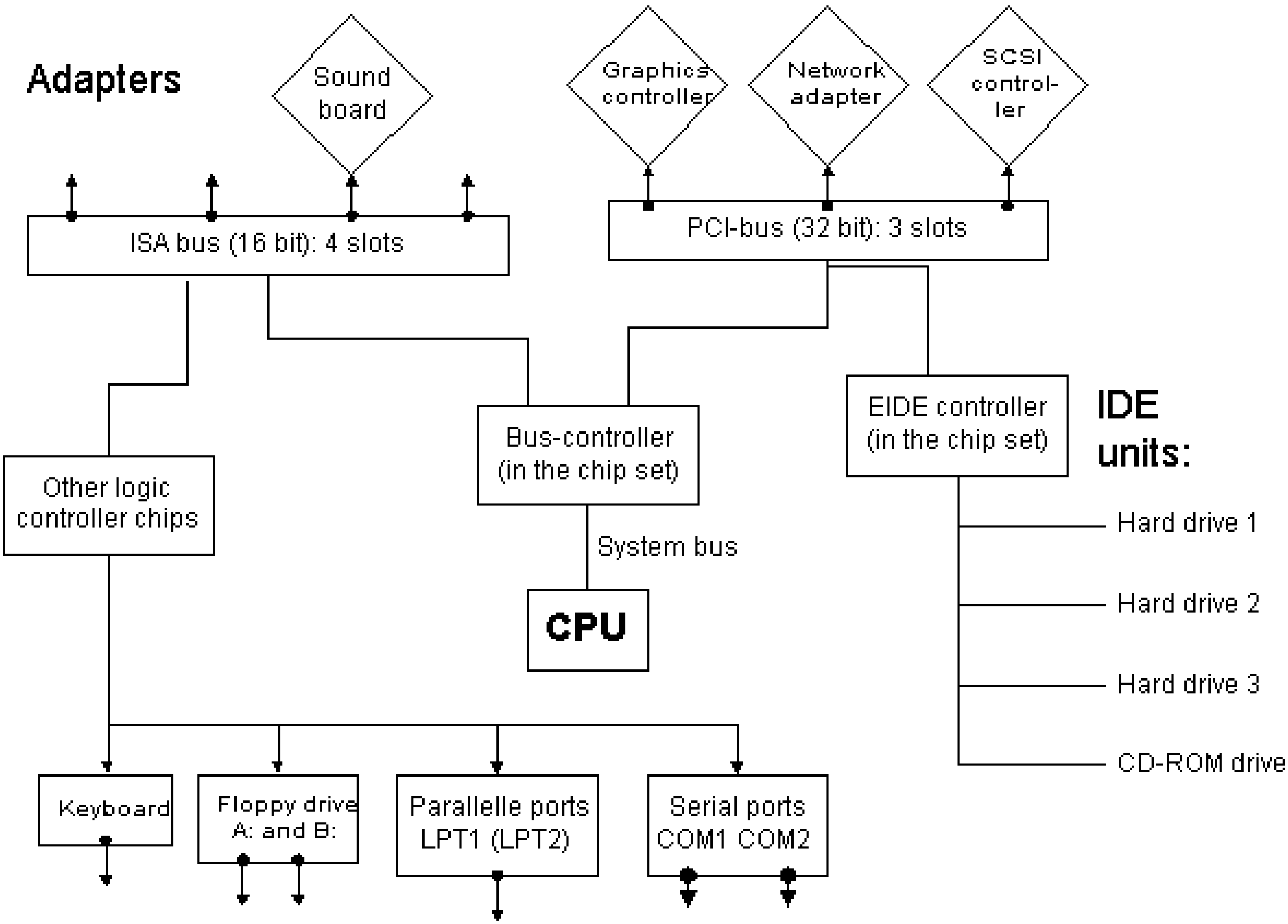


The data flow on the system board

On the system board, you will find the CPU, which is the "brain" of the PC and the *busses*. The busses are the nerve system of of system board. They connect the CPU to all the other components. There are at least three busses, which you can see below. You can read more about those on the following pages.

- **The busses** are the PC's expressways. They are "wires" on the circuit board, which transmit data between different components. One "wire" can move one bit at a time**.**

In the following text, we start from a modern Pentium board. We will look at busses, chip sets and CPU's. Here is an illustration of the system board "logic," which you can print out.



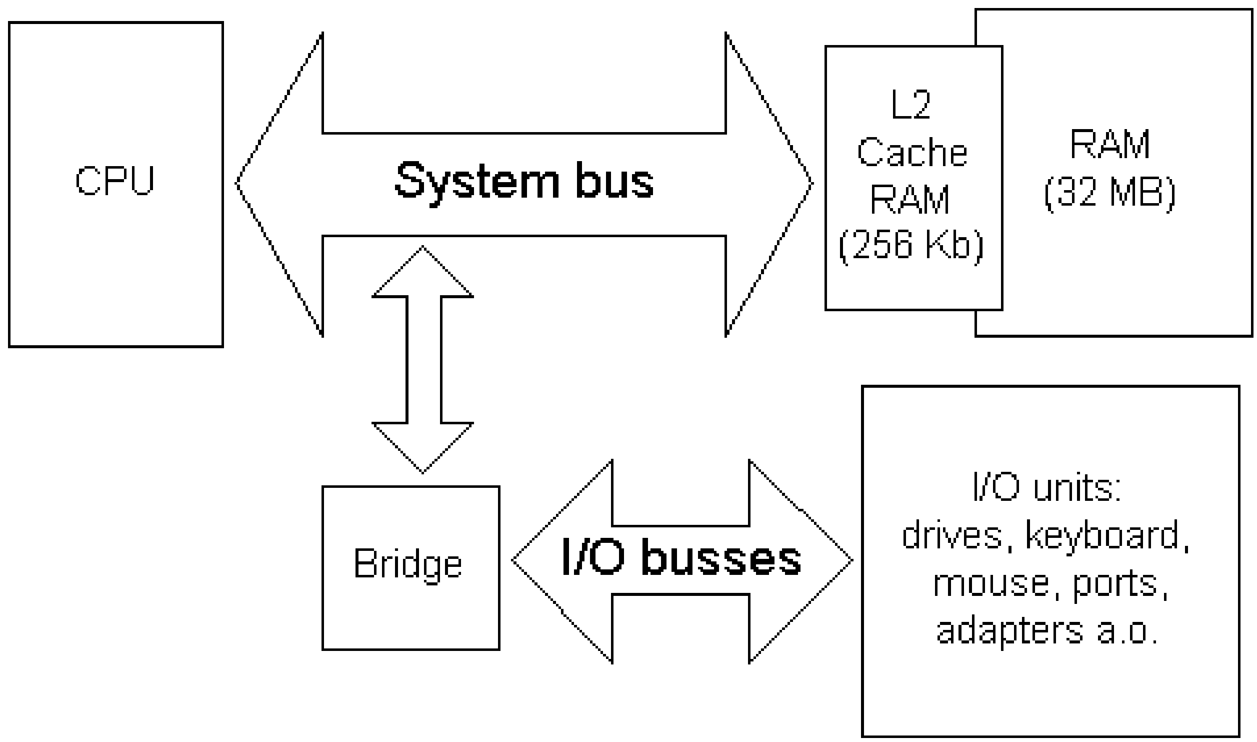
Here, you can switch to read about what data really are. Or click to move to the page about the CPU. Or read more about data, and how they are saved in the PC file systems. Or just continue here.

Introduction to the PC busses

The PC receives and sends its data from and to *busses*. They can be divided into:

- **The system bus, which connects the CPU with RAM**
- **I/O busses, which connect the CPU with other components.**

The point is, that the *system bus* is the central bus. Actually, it connects to the I/O busses, as you can see in this illustration:



You see the central system bus, which connects the CPU with RAM. A *bridge* connects the I/O busses with the system bus and on to RAM. The bridge is part of the PC chip set, which will be covered in module 2c.

---

3 different I/O busses

The I/O busses move data. They connect all I/O devices with the CPU and RAM. I/O devices are those components, which can receive or send data (disk drives, monitor, keyboard, etc.). In a modern Pentium driven PC, there are two or three different I/O busses:

- **The ISA bus, which is oldest, simplest, and slowest bus.**
- **The PCI bus, which is the fastest and most powerful bus.**
- **The USB bus, which is the newest bus. It may in the long run replace the ISA bus.**

The three I/O busses will be described later. Here, we will take a closer look at the PC's fundamental bus, which the others are branches from:

The system bus

The system bus connects the CPU with RAM and maybe a buffer memory (L2-cache). The system bus is the central bus. Other busses branch off from it.

The system bus is on the system board. It is designed to match a specific type of CPU. Processor technology determines dimensioning of the system bus. At the same time, it has taken much technological development to speed up "traffic" on the system board. The faster the system bus gets, the faster the remainder of the electronic components must be..

The following three tables show different CPU's and their system busses:

Older CPU's	System bus width	System bus speed
8088	8 bit	4,77 MHz
8086	16 bit	8 MHz
80286-12	16 bit	12 MHz
80386SX-16	16 bit	16 MHz
80386DX-25	32 bit	25 MHz

We see, that system bus speed follows the CPU's speed limitation. First at the fourth generation CPU 80486DX2-50 are doubled clock speeds utilized. That gives the CPU a higher *internal* clock frequency. The *external* clock frequency, used in the system bus, is only half of the internal frequency:

CPU's in the 80486 family	System bus width	System bus speed
80486SX-25	32 bit	25 MHz
80486DX-33	32 bit	33 MHz
80486DX2-50	32 bit	25 MHz

80486DX-50	32 bit	50 MHz
80486DX2-66	32 bit	33 MHz
80486DX4-120	32 bit	40 MHz
5X86-133	32 bit	33 MHz

---

66 MHz bus

All the first Pentium based computers ran at 60 or 66 MHz on the system bus, which is 64 bit wide:

CPU's in the Pentium family	System bus width	System bus speed
Intel P60	64 bit	60 MHz
Intel P100	64 bit	66 MHz
Cyrix 6X86 P133+	64 bit	55 MHz
AMD K5-133	64 bit	66 MHz
Intel P150	64 bit	60 MHz
Intel P166	64 bit	66 MHz
Cyrix 6X86 P166+	64 bit	66 MHz

Pentium Pro 200	**64 bit**	**66 MHz**
Cyrix 6X86 P200+	**64 bit**	**75 MHz**

---

100 MHz bus

The speed of the system bus will increase in 1998. Using PC100 SDRAM a speed of 100 MHz is well proven and later the use of RDRAM will give us much higher speeds.

However the raise from 66 MHZ to 100 MHZ will have the greatest impact on socket7 CPU's and boards. In the Pentium-II modules 70-80% of the traffic is inside the SEC-module, holding both L1 and L2 cache. And the module has it own speed independent of the system bus. But with the K6 the increase of system bus speed will give a vastly improved performance since the traffic between L12 and L2 cache crosses the system bus.

CPU	Chip set	System bus speed	CPU speed
Intel Pentium II	**82440BX**	**100 MHz**	**350, 400, 450 MHZ**
AMD K6+ **AMD K6-3D**	**?**	**100 MHz**	**250, 300, 400 MHZ**
Intel Pentium III	**82450NX**	**100 MHz**	**450, 500 MHZ**

Read more about I/O busses in module 2c

Read more about the system boards chip set in module 2d

Read more about RAM in module 2e

To overview

Click & Learn. Module 2c. WWW.MKDATA.DK

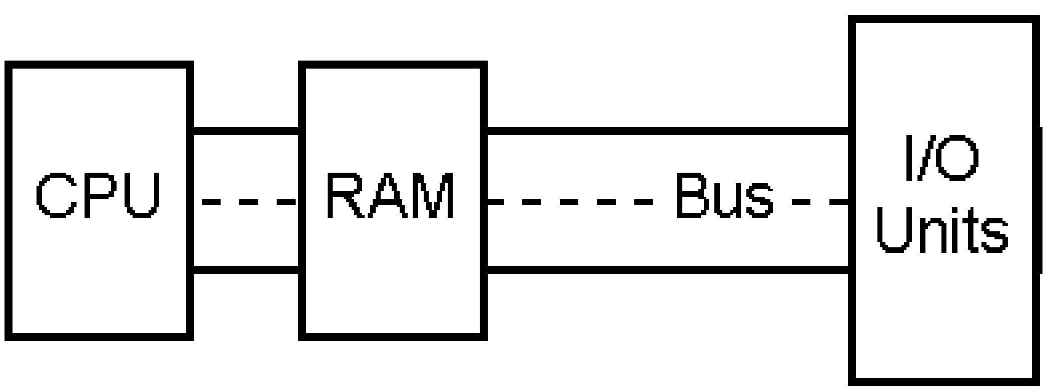
About the I/O busses

On these pages, you can read about the important system bus derivatives, the different I/O busses:

Introduction to the I/O busses

We have seen before, that the PC's busses are the fundamental data "highways" on the system board. The "first" bus is the *system bus*, which connects the CPU with RAM. It is also called the local bus. Its speed and width depends on the type CPU installed on the system board. Typically, the system bus will be 64 bits wide and run at 66 MHz. That high speed creates some electrical noise and other problems. Therefore, the speed must be reduced when we need to reach the expansion cards and some other components. Very few expansion cards can operate at more than 40 MHZ. Then the electronics shut down. The chips can just not react faster. Therefore, the modern PC has additional busses.

However, the first PC's had only one bus, which was common for the CPU, RAM and I/O components:



The older first and second generation CPU's ran at relatively low clock frequencies, and all system components could keep up with those speeds. Among other things, that allowed additional RAM to be installed in expansion slots in the PC, by installing an adapter in a vacant expansion slot. An adapter, where RAM was mounted. That would be unthinkable today.

First in 1987, Compaq figured how to separate system bus from I/O bus, so they could run at different speeds.. This multi-bus architecture has been industry standard ever since. Modern PC's also have more than one I/O bus.

What does an I/O bus do?

I/O busses connect the CPU to all other components, except RAM. Data are moved on the busses from one component to another, and data from other components to the CPU and RAM. The I/O busses differ from the system bus in speed. Their speed will always be lower than the system bus speed. Over the years, different I/O busses have been developed. On modern Pentium PC's, you will find at least two significant busses, and one less significant:

- **The ISA bus**, which is an older low speed bus.
- **The PCI bus**, which is a new high speed bus.
- **The USB bus** (*Universal Serial Bus)***,** which is a new low speed bus.

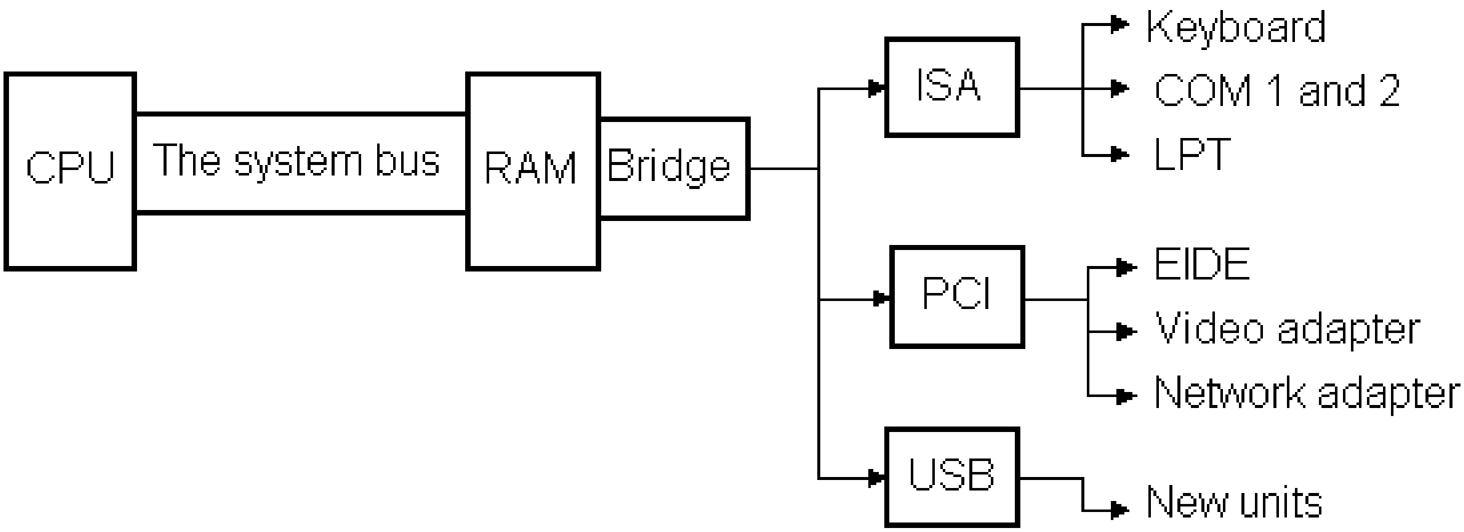
As mentioned earlier, I/O busses are really derivatives from the system bus. On the system board it ends in a controller chip, which forms a bridge to the two other busses.

The essential in modern PC's is fast busses. Let us compare the two primary I/O busses:

Bus	Transmission time	Data volume per transmission
ISA	375 ns	16 bit
PCI	30 ns	32 bit

Clearly, there is a vast difference between the capacity of the two busses.

All in all, the busses have a very central placement in the PC's data exchange. Actually, all components except the CPU communicate with each other and with RAM via the different I/O busses. Here you see a demonstration of this logic:



The physical aspects of the I/O busses

Physically, the I/O bus consists of one or more tracks on the printed circuit board. These tracks are used as:

- Data tracks, which each can move one bit at a time
- Address tracks, which identify where data should be sent to
- Other tracks for clock ticks, voltage, verification signals, etc.

Wen data are sent on the bus, they must be supplied with a *receiver*. Therefore, each device on the bus has an address. Similarly, the RAM is divided in sections, each having its address. Prior to sending data, a number is sent on the address track, to identify where the data should be sent to.

The bus width. The number of data tracks determine the data transfer capacity. The ISA bus is slow, partly because it only has 16 data tracks. The modern PC's send 32 bits per clock tick. On the ISA bus, 32 bits must be divided in two packages of 16 bits. This delays the data transfer. Another I/O bus concept is *wait states.*

Wait states are small pauses. If an ISA adapter can not keep up with the incoming data flow, its controller sends wait states to the CPU. Those are signals to the CPU to "hold on for a sec." A wait state is a wasted clock tick. The CPU skips a clock tick, when not occupied. Thus the old and slow ISA adapter can significantly reduce the operating speed of a modern computer.

Another aspect is the IRQ signals, which the components use to attract attention from the CPU. That and the concepts *DMA* and *bus mastering,* are described in module 5, which deals with adapters.

Technical and historical background for the I/O busses

In modern PC's you only find the PCI and ISA busses (besides USB, which we do not know much about yet). But, over the years, there have been other busses. Here is a diagram of the various I/O busses. Then comes a more detailed description of each of the busses:

Bus	Year	Bus width	Bus speed	Max. throughput rate (theoretical)
PC and XT	1980-82	8 bit	4,77 and 6 MHZ	4-6 MBps
ISA (AT) Simple bus, cheap. non-intelligent.	1984	16 bit	8 MHZ	8 MBps
MCA Advanced, intelligent bus. Used almost exclusively by IBM.	1987	32 bit	10 MHZ	40 MBps
EISA Advanced, intelligent bus. Used by net servers.	1988	32 bit	8 MHZ	32 MBps
VL Simple high speed bus Genuine local bus, synchronous with the CPU. Used in 486's.	1993	32 bit	Synchronous with the CPU: 25, 33 or 40 MHZ	100-160 MBps

PCI Intelligent, advanced high speed bus. Standard with Pentium.	1993	32 bit	Asynchronous: 33 MHZ	132 MBps

The ISA bus

Since about 1984, standard bus for PC I/O functions has been named ISA (*Industry Standard Architecture*). It is still used in all PC's to maintain backwards compatibility. In that way modern PC's can accept expansion cards of the old ISA type.

- ISA was an improvement over the original IBM XT bus, which was only 8 bit wide. IBM's trademark is *AT bus*. Usually, it is just referred to as ISA bus.
- ISA is 16 bit wide and runs at a maximum of 8 MHz. However, it requires 2-3 clock ticks to move 16 bits of data.
- The ISA bus has an theoretical transmission capacity of about 8 MBps. However, the actual speed does not exceed 1-2 MBps, and it soon became too slow.

The ISA bus has two "faces" in the modern PC:

- **The internal ISA bus**, which is used on the simple ports, like keyboard and serial or parallel ports.
- **As external expansion bus**, which can be connected with 16 bit ISA adapters. ISA-slots is especially used with the common 16 bit sound boards.

The problem with the ISA bus is partly its narrow width. It can not transfer enough bits at a time, and partly its lack of "intelligence."

The lack of intelligence means that the CPU controls the data transfer across the bus. The CPU can not start a new assignment, until the transfer is completed. You can observe that, when your PC communicates with the floppy drive (the LED on the floppy drive drive goes on), while the rest of the PC is waiting. That is the fault of the ISA bus.

The ISA bus can be a tease, when you install new expansion cards (for example a sound card). Many of these problems derive from the tuning of IRQ and DMA, which must be done manually on the old ISA-bus. Every component occupies a specific IRQ and possibly a DMA channel. That can create conflict with existing components. Read module 5 about expansion cards and these problems.

MCA, EISA and VLB

In the 80's, a demand developed for busses more powerful than the ISA. IBM developed the MCA bus and Compaq and others responded with the EISA bus. None of those were particularly fast, and they never became particularly successful outside the server market.

MCA

IBM's top of the line bus from 1987 is named *Micro Channel Architecture*. Contrary to ISA, MCA is patented, and IBM demanded high royalty fees, when other PC manufacturers wanted to use it. Thus it never became a great success, despite its advanced design. It was a classic example of poor marketing strategy.

The MCA bus is 32 bit wide and "intelligent." The cards configure themselves with respect to IRQ. Thus, they can be installed without adjustments of jumper switches or other features. The MCA bus is also relatively fast with transfer rates of up to 40 MBps in 32 bit mode at 10,33 MHz. MCA requires special adapters. There have never been too many adapters developed, since this bus is by and large used only in IBM's own PC's.

EISA

EISA is a bus from 1988-89. It is designed by the "Gang of Nine:" the companies AST, Compaq, Epson, Hewlett-Packard, NEC, Olivetti, Tandy, Wyse and Zenith. It came in response to IBM's patented MCA bus.

EISA is intelligent with bus mastering, divided interrupts and self configuration. It is 32 bit wide, and runs at 8 MHZ. But, like the MCA, it did not have great success. EISA is compatible with ISA in the sense that ISA adapters can be installed in EISA slots. The EISA bus is still used in many servers.

Vesa Local Bus

This Bus called VLB for short. It is an inexpensive and simple technology. This bus only achieved status as an interim phenomenon (in 1993-94). VLB was widely used on 486 system boards, where the system bus runs at 33 MHZ. VLB runs directly with the system bus. Therefore, data transfer is at CPU speed, synchronous and in width. The problem with VLB was compatibility. Adapters and system system boards would not always work together. Vesa is an organization with about 120 members, mostly monitor and graphics card manufacturers. Therefore, most VLB cards were video cards.

The PCI bus

The PCI is the nineties high speed bus for PC's. PCI stands for *Peripheral Component Interconnect*. This bus is made by Intel. It is used today in all PC's and other computers.

The PCI is actually 32 bit wide, but in practice it functions like a 64 bit bus. It has a maximum transmission capacity of 132 MBps. PCI is the most advanced (and, most expensive) bus. According to the specifications - not in practice, it can have up to 8 units with a speed up to 200 MHZ . The bus is processor independent. Therefore, it can be used with all 32 or 64-bit processors, and it is found also on other computers than PC's. It is also compatible with the ISA bus, in that the PCI bus can react on ISA bus signals, create the same IRQ's, etc.

The PCI bus is *buffered* in relation to the CPU and the peripheral components. This means, that the CPU can deliver its data to the buffer, and then proceed with other tasks. The bus handles the further transmission in its own tempo. Conversely, the PCI adapters can also transmit data to the buffer, regardless of whether the CPU is free to process then. They are placed in a queue, until the system bus can forward them to the CPU. Under optimal conditions, the PCI bus transmits 32 bits per clock tick. Sometimes, it requires two clock ticks.

Because of this, the peripheral PCI units operate *asynchronous* . Therefore, the PCI (contrary to the VL bus) is not a local bus in a strict sense. Finally, the PCI bus is intelligent relative to the peripheral components, in that Plug and Play is included in the PCI specifications. All adapter cards for the PCI configure themselves. Plug and Play is abbreviated PnP.

On modern system boards, the PCI bus (like ISA) has two "faces:"

- **Internal PCI bus**, which connects to EIDE channels on the system board.
- **The PCI expansion bus**, which typically has 3-4 slots for PCI adapters.

The PCI bus is continuously being developed further. There is a PCI Special Interest Group, consisting of the most significant companies (Intel, IBM, Apple, and others), which coordinate and standardize the development. It has been announced, that PCI should be developed with a higher bus speed (66 MHZ) and greater width (64 bit). At the moment, it looks like alternative busses will be marketed. An example is the high speed AGP video bus (*Accelerated Graphics Port*).

Read more about chip sets on the system board in module 2d

Read more about RAM in module 2e

Read module 5a about expansion cards, where we evaluate the I/O busses from the port side.

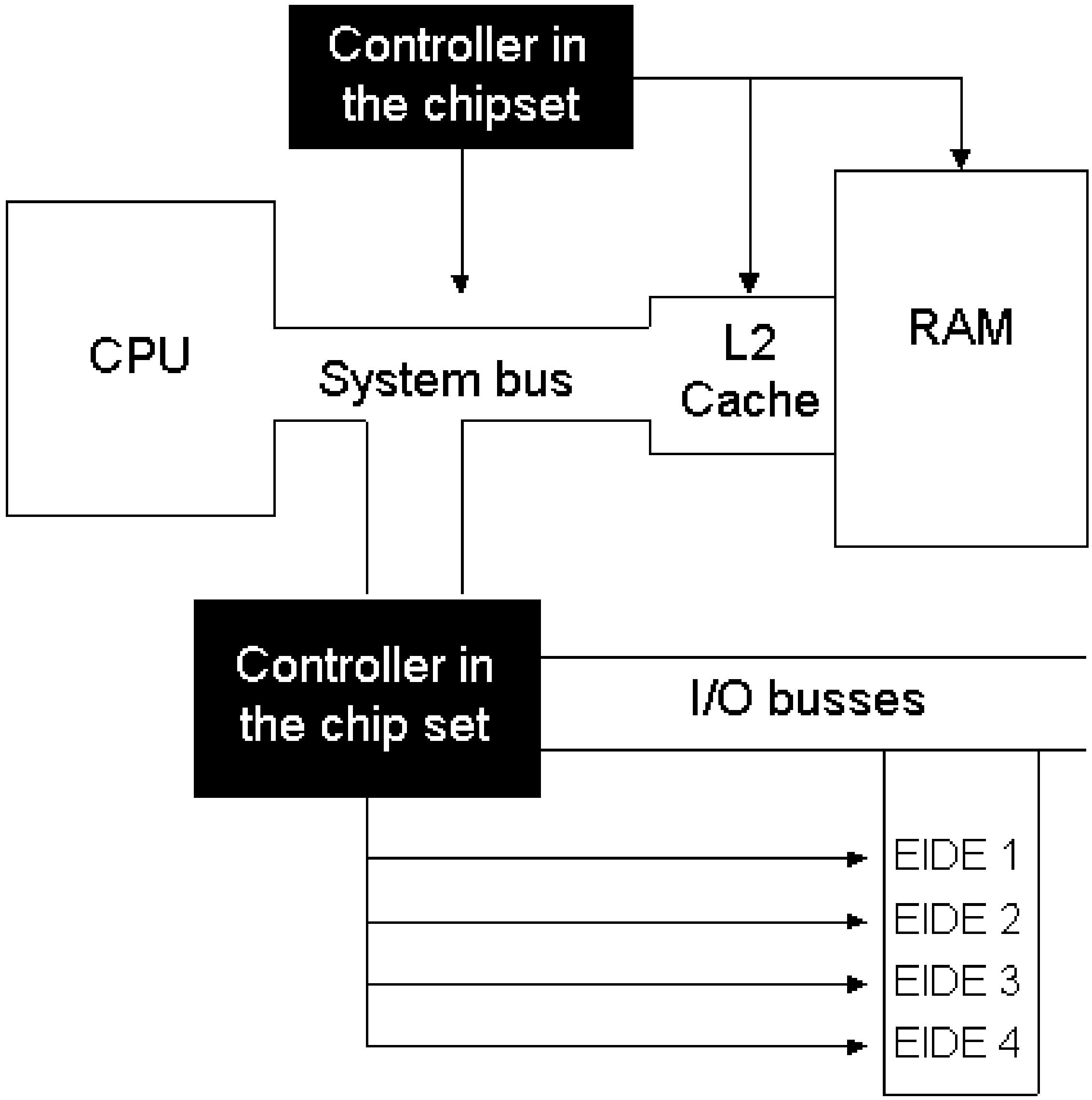
To overview

Click & Learn. Module 2d. WWW.MKDATA.DK

Chip sets

What is a chip set?

When we speak about busses and system boards, we are also speaking about *chip sets*. The chip sets are a bunch of intelligent controller chips, which are on any system board. They are closely tied to the CPU, in that they control the busses around the CPU. Without the chip sets, neither RAM or I/O busses could function together with the CPU:



New technologies - new chip set

Therefore, the chip sets are quite central components on the system boards. When new technological features are introduced (and this happens continuously) they are often accompanied by new chip sets. The new chip sets often enable:

- **Higher speed on one or more busses**
- **Utilization of new facilities (new RAM types, new busses, improved EIDE, etc.)**

There are several suppliers of Pentium chip sets:

- **Intel**
- **SIS**
- **Opti**
- **Via**

- **AMD**

Intel has hitherto been the leader in supplying chip sets to the Pentium system board. Therefore, let us just mention their chip sets, which have astronomical names.

The Neptune chip set (82434NX) was introduced in June 1994. It replaced the Mercury set (82434LX). In both chip sets, there were problems with the PCI bus. In January 1995 Intel introduced the first Triton , where everything worked. This chip set supports some new features: it supports EDO RAM, and it offers bus master integrated EIDE control and NSP (*Native Signal Processing* - one of the many new creations, which was soon forgotten).

Triton first and second

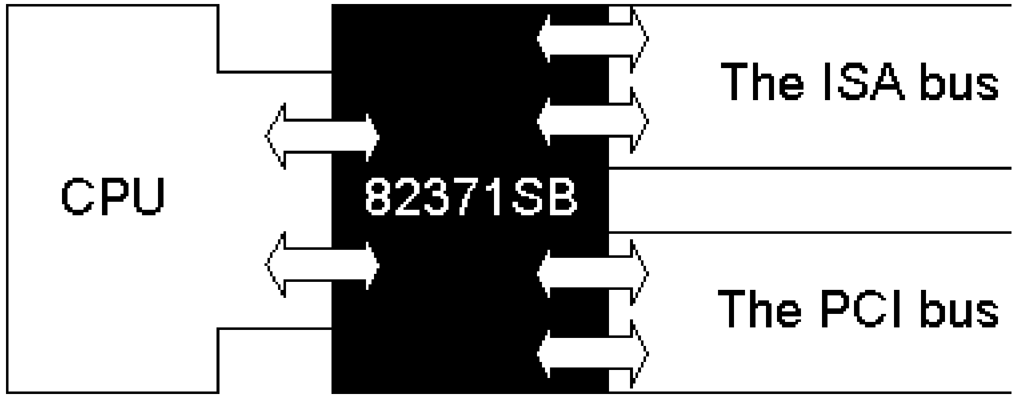
82430FX from late 1995 was Intel's next chip set and the first Triton. In February 1996 the second generation of Triton arrived. Two new chip sets were introduced: The 82430VX and 82430HX. The last (HX) was the fastest one.

The two sets are similar, yet different. 430HX consists of two chips. It is designed for the more professional PC's. 430VX consists of four chips. Its cost is slightly lower than HX. It is aimed at the home use PC market. Let us look at the contents of each chip set:

Chip set	Contents
82430HX	82439HX System Controller (TXC) + 82371SB PCI ISA IDE Accelerator
82430VX	82437VX System Controller (TVX) + 2 stk. 82438VX Data Path Units (TDX) + 82371SB PCI ISA IDE Accelerator

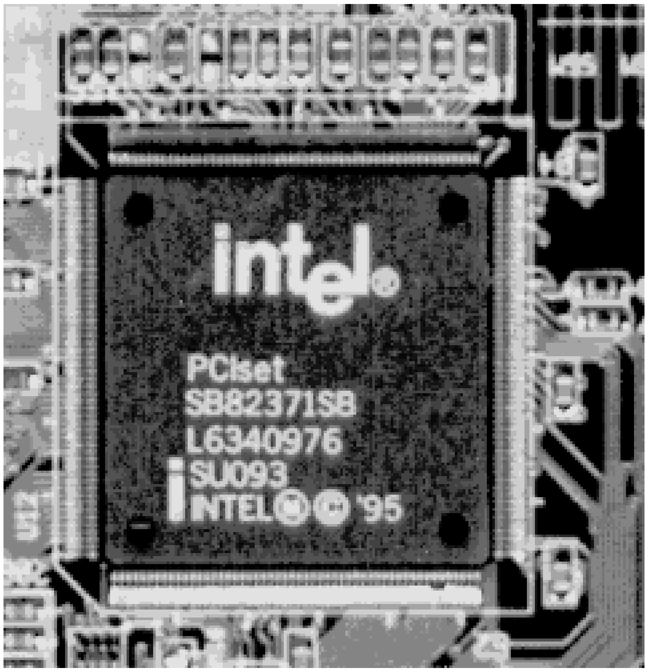
Common to both chip sets is **82371SB**, which is a "PCI ISA IDE accelerator chip". It is also called PIIX3, which some may recognize from the Windows 95 device driver, which comes with the ASUS T2P4 board.

The chip makes a bridge between the CPU, ISA and PCI bus. The new is, that it permits concurrent activity in all three locations, thus a new form of multitasking. This is significant for daily use. All data exchange to and from I/O units cross this intersection, which now has achieved greater width:

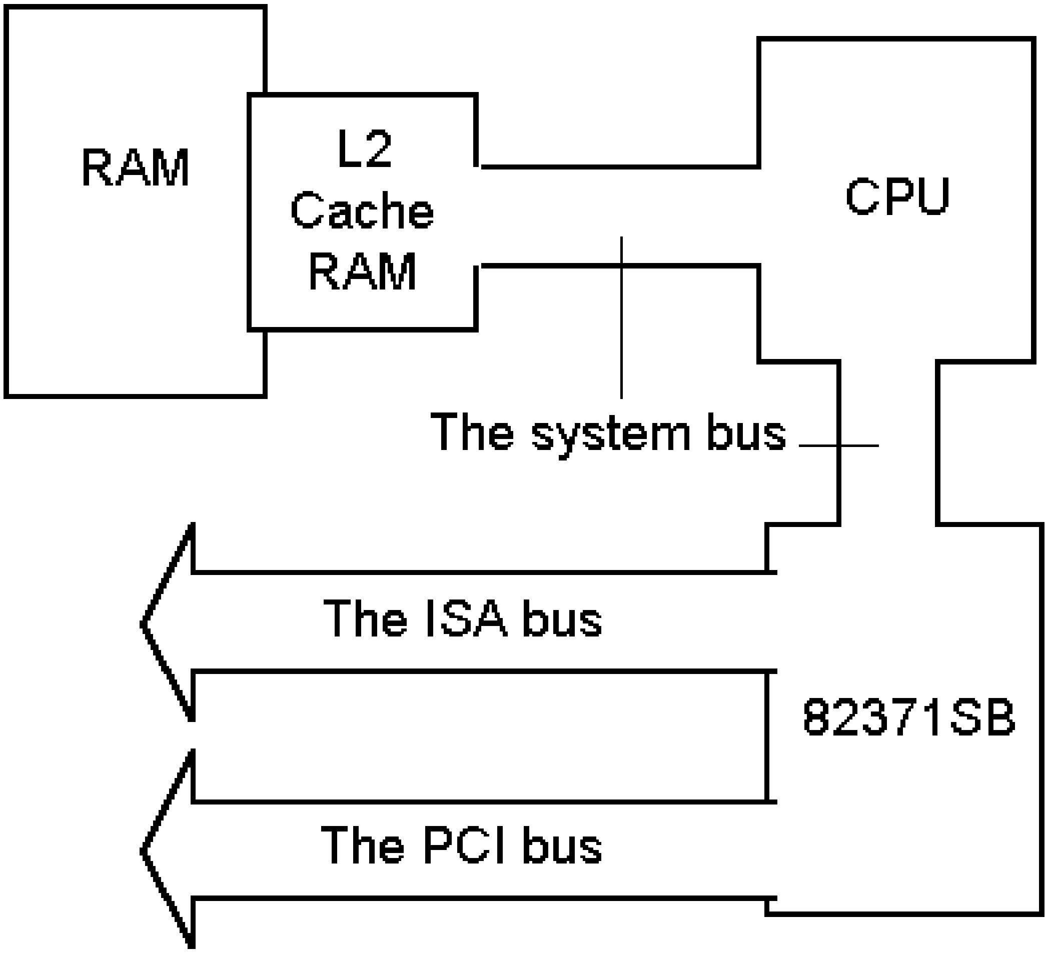


.

New in the chip is also a host function for the USB (*Useless Serial Bus)*, which we have not seen much use of. Finally, the chip includes EIDE Bus Master control. That means, that EIDE components like hard disks, to some extent can deliver their data directly to RAM without taking up CPU time.



Above, you see the 82371SB chip and below, again, its placement relative to CPU and busses:



It is generally accepted, that the HX set yields the best performance of the two chip sets described. But the VX set has two other facilities to offer: Capability for SMBA (*Shared Memory Buffer Architecture*). That means among other things, that you can integrate the video card on the system board with 1 or 2 MB standard RAM, from the working RAM. A technology, which is used only in the lowest cost PC's.

Also, the VX set also supports the fast RAM type SD-RAM. HX does not. The VX set can control up to 128 MB RAM, but it cannot cache above 64 MB RAM.

HX controls 512 MB RAM and is the only Intel Pentium chip set to cache above 64 MB RAM.

The VX and HX chip sets are out. They are replaced by the TX chip set.

Intel TX chip set - AGP and Ultra DMA

The newest chip set to Pentium processors are named 82430TX and 82430LX. Both support SD-RAM.

New features are AGP, which is a new high speed graphics bus (*Accelerated Graphics Port*).

The AGP-bus runs at 66 MHZ - twice the speed of the PCI bus. This gives new power to the video system, and it frees the PCI bus from all the heavy video work. The AGP adapters also extend their memory using parts of the mainboard RAM as a bit map cache.

ATA-33 permits EIDE hard disks to transfer at up to 33 MBps - a data volume which no hard disk can deliver. This improved EIDE standard is also marketed under the name Ultra DMA. Tests show that Ultra DMA results in a speed increase of 25-75 percent over the traditional EIDE PIO mode 4. Ultra DMA is the new EIDE standard.

Chip set	Chips included
82430TX	82439TX System Controller (TXC) 82371AB PCI ISA IDE Accelerator

The TX set is an update and improvement of the VX set. Relative to this, the TX firstly supports SD RAM and Ultra DMA hard disks. Two important technologies. But the TX-set cannot cache above 64 MB RAM, and that is a problem.

Photos taken with Canon Powershot 600

Soyo ETEQ

Since Intel does not develop new chip sets for Socket7 main boards, it is interesting to follow companies like Soyo. This the worlds 7th biggest main board manufactor has deloped their own ETEQ 8236638AT/6629 AGP chipset (ETEQ 6638 for short) which gives new performance using a bus connecting the CPU with the L2 cache at 100 MHZ. The RAM works at only 66 MHZ.

No more informations available - but tests shows that their board SY 5EH5/M (when do taiwanese manufactors start naming their products E525 or something to remember?) performs very well with AMD K6-300 at 3 X 100 MHZ.

VIA Apollo MVP3

This chip set for Socket7 main boards promises very good performance with its 100 MHZ bus connecting the CPU with the L2 cache and possible PC100 SDRAM. Other features include:

- AGP graphics
- Supports up to 1GB 64Mbit FP/EDO/SDRAM/ PC100
- 2MB L2 cache support
- Virtual Clock Synchronization (VCS) for optimal timing.

The Apollo MVP3 chipset consists of a VT82C598AT system controller (476 pin BGA, the "north bridge") and a VT82C586B PCI to ISA bridge (208 pin PQFP, the "south bridge"):

See VIA's own data

Also SIS produces new chip set for Socket7 boards.

Chip sets for Pentium Pro and Pentium II

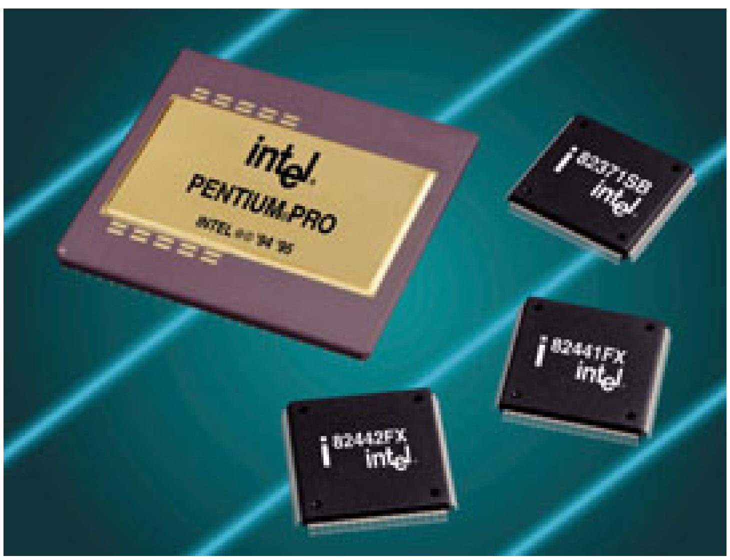
The 6th. generation CPU's, Pentium Pro and Pentium II, have their own chip sets. Let us review them chronologically:

82450GX

This chip set came out in 1995, It is supporting quad CPU configurations. That is PC's with 4 Pentium Pros.

82440FX - Natoma

This is Intel's most widely used chip set for 6th. generation CPU's. This chip set can handle 2 CPU's on the same system board. 440FX consists of four controllers. As for features, it is like the 82430HX set. Common for these chip sets is the 82371SB PCI-ISA accelerator, which provides good performance on the I/O busses.



This chip set is good and fast. However, it does not support neither SDRAM, Ultra DMA, or AGP. These features are found in the following chip set 82440LX.

82440LX

440LX is from August 1997. The new is in AGP and SD-RAM. Thus, this chip set is equivalent to 430XT.

82440BX

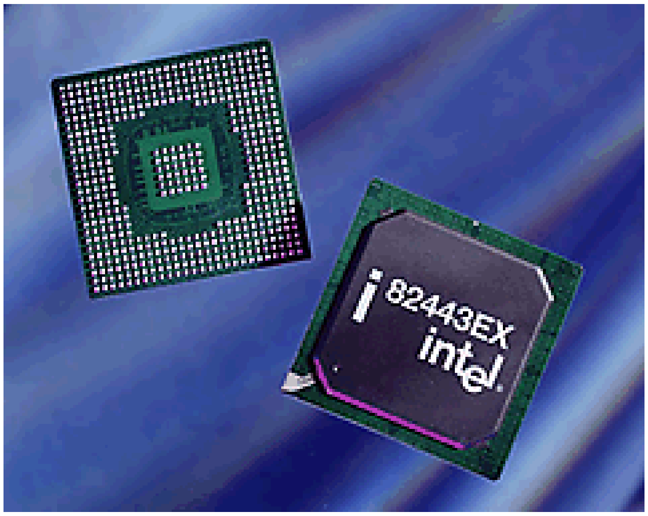
440BX was released on April the 15th 1998. The chip set contains the 82443BX Host Bridge and the 82371EB PIIX4E:

The system bus speed is increased to 100 MHZ. This chip set is is designed for the new Pentium II chips, which will run at 350, 400, and later 450 MHZ. The 100 MHZ system bus speed is multiplied with clock factors of 3.5, 4, and 4.5.

The chip set promises new and better bandwidth for PCI and AGP using a new Quad Port technology, which is not explained anywhere. It was expected that the BX chip set should support the IEEE 1394 bus (FireWire) but it does not yet.

82440EX

The EX chip set is a discount version of the LX set. The chip set only supports DIMM sockets with up to only 256 MB RAM and just three PCI slots. To be used with the inexpensive Celeron cartridges.



82450NX

Intel will introduce completely new versions of Pentium II using the new Slot 2.

450NX is the first chip set for that allowing 4 or 8 CPU's on the main board using up to 4GB RAM. An other new feature is the 66 MHZ PCI bus. This chip set is for servers.

Carmel

The BMW/Volkswagen version of the 82450NX is codenamed Carmel. Here we find support for 4X AGP, RDRAM, UDMA66 as well as the 66 MHZ PCI bus.

Other new chip sets are:

- 82440GX, which is a BX set supporting up to 2 GB RAM
- 133 MHZ Front Side Bus for RDRAM.
- Camino for the Katmai CPU, being much similar to Carmel from what I read.
- Whitney, which is a LX set with integrated i740 graphics controller.

If you want to read more about these and other chip sets, look for the excellent German/American web site Toms Hardware Guide. Here, you will find all about these subjects.

Read more about RAM in module 2e

Read about the Pentiums ao. in module 3c

To overview.

Click & Learn. Module 2e. WWW.MKDATA.DK

About RAM

This page should be read together with modules 2a, 2b, 2c, and 2d, which deal with system board, system bus, I/O bus and chip sets. When we talk about system board and chip sets, we can not ignore RAM. Warning: RAM and RAM chips is a very complicated, technical subject area. I can in no way give a complete, comprehensive description of this subject. But, here you can read a little about:

What is RAM?

RAM is our working memory storage. All the data, which the PC uses and works with during operation, are stored here. Data are stored on drives, typically the hard drive. However, for the CPU to work with those data, they must be read into the working memory storage, which is made up of RAM chips. To examine RAM, we need to look at the following:

- **RAM types (FPM, EDO, ECC, and SD RAM)**
- **RAM modules (SIMM and DIMM) in different versions**
- **RAM and the system bus**

First, let us look back in time. Not too many years ago, Bill Gates said, that with 1 MB RAM, we had a memory capacity, which would never be fully utilized. That turned out to be untrue.

Historical review

Back in the 80's, PC's were equipped with RAM in quantities of 64 KB, 256 KB, 512 KB and finally 1 MB. Think of a home computer like Commodore 64. It had 64 KB RAM, and it worked fine.

Around 1990, advanced operating systems, like Windows, appeared on the market, That started the RAM race. The PC needed more and more RAM. That worked fine with the 386 processor, which could address larger amount of RAM. The first Windows operated PC's could address 2 MB RAM, but 4 MB soon became the standard. The race has continued through the 90's, as RAM prices have dropped dramatically.

Today. it would be foolish to consider less than 32 MB RAM in a PC. Many have much more. I have 128 MB, and that is in no way too much. Especially, if you use Windows 95, it is important with plenty of RAM. Click here to read about the swap file and RAM considerations.

RAM types

The traditional RAM type is DRAM (*dynamic* RAM). The other type is SRAM (*static* RAM). SRAM continues to remember its content, while DRAM must be refreshed every few milli seconds. DRAM consists of micro capacitors, while SRAM consists of off/on switches. Therefore, SRAM can respond much faster than DRAM. SRAM can be made with a *rise time* as short as 4 ns. It is used in different versions in L2 cache RAM (for example *pipe line Burst Cache SRAM*).

DRAM is by far the cheapest to build. Newer and faster DRAM types are developed continuously. Currently, there are at least four types:

- **FPM (*Fast Page Mode*)**
- **ECC (*Error Correcting Code*)**
- **EDO (*Extended Data Output*)**
- **SDRAM (*Synchron Data RAM*)**

A brief explanation of DRAM types:

FPM was the traditional RAM for PC's, before the EDO was introduced. It is mounted in SIMM modules of 2, 4, 8, 16, or 32 MB. Typically, it is found in 60 ns or 70 ns versions. 60 ns is the fastest and the one to use. You cannot mix different speeds on the same Pentium system board.

EDO is an improvement of FPM RAM. Data are read faster. By switching from FPM to EDO, one can expect a performance improvement of 2 to 5 percent. EDO RAM are usually sold in 60 ns versions. A 50 ns version is available at higher cost.

ECC RAM is a special error correcting RAM type. It is especially used in servers.

SDRAM is the newest RAM type for PC's. It comes only in 64 bit modules (long 168 pin DIMM's). SDRAM has a rise time of only 8-12 ns. The performance improvement over EDO RAM is only 5 percent running at 66 MHZ, but at 100 MHZ it will prove a lot better.

RAMBUS (RDRAM) is a future RAM type. Intel and others have great expectations from this type.

---

8 or 9 bits per byte?

Normally you figure 8 bits to one byte. For many years, a ninth bit has been added as *parity* bit in the RAM blocks to verify correct transmission. That way you have to transmit 9 bits, to store 8 bits in the old 30 pin RAM chips. And it takes 36 bits to store 32 bits in the larger 72 pin chips.

which increases the cost of the RAM chip by about 12%.

If your system board requires 36 bit modules, you must respect that. Fortunately, most system boards accepts 32 bit modules, so this creates no problems.

RAM and system board

You can not freely install your desired RAM type. RAM is controlled by the chip set on the system board, so you must install a type, which matches your system board. Furthermore, RAM chips come in different sizes, which must match the system board.

On modern system boards, RAM is installed on SIMM or DIMM modules. Before, small individual DRAM's were used. There was usually room for 36 small chips on the system board. That made it cumbersome to install new RAM. Then, someone figured out to install RAM chips on cards, which are easily installed. First came the SIPP modules. They had multiple pins, which fit in the system board. Since came the SIMM modules. They are mounted on a card, which has an edge connector. They fit in sockets on the system board, and anyone can install them.

RAM speeds

RAM speed is measured in ns (*nano seconds*). The fewer ns, the faster is the RAM. Years ago, RAM came in 120, 100 and 80 ns. Today, we are talking about 60 ns and faster.

It becomes complicated to describe the relationship between RAM speed and the ability of the system bus to utilize fast RAM. I will gloss over that. But here is a table which illustrates RAM speed, relative to clock speed:

Clock speed	Time per clock tick
20 MHZ	50 ns
25 MHZ	40 ns
33 MHZ	30 ns
50 MHZ	20 ns
66 MHZ	15 ns
100 MHZ	10 ns

Peak Bandwidth

Here you see the maximal peak bandwidth of the three well-known RAM-types. The figures illustrates the absolutely maximal transfer from RAM to the L2-cache - in peaks, not a continuously transfer.

RAM-type	Max. peak bandwidth
FPM	176 MB/sec
EDO	264 MB/sec
SD	528 MB/sec

SIMM's

SIMM modules were first made in 8 bit editions. They were small cards with 1, 2 or 4 MB RAM. They were connected to the system board with a 30 pin edge connector. The modules were 8 bit wide. This meant that 16 bit processors (286 and 386SX) needed 2 SIMM's in a pair. Thus, there was room for two modules in what is called a *bank*.

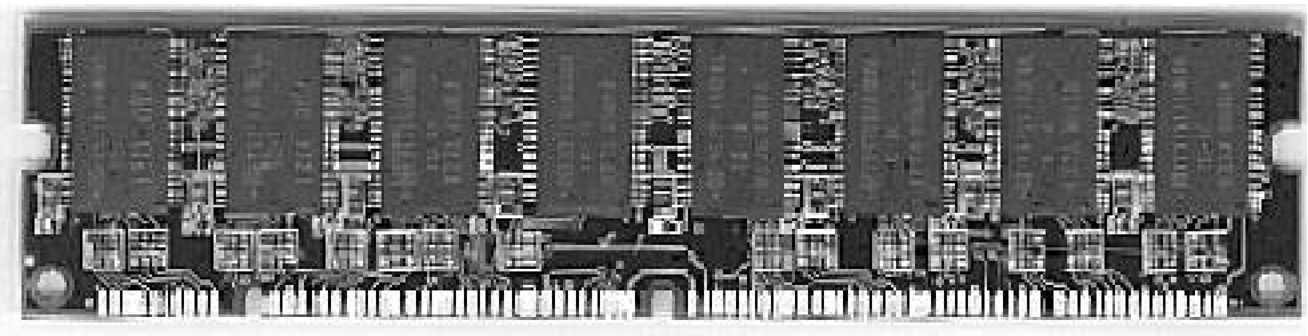
32 bit processors (386DX and 486) need 4 of the small 8 bit SIMM's in a bank, since their banks are 32 bit wide. So, on a typical 1st generation 486 system board, you could install 4 X 1 MB, 4 X 2 MB, or 4 X 4 MB in each bank. If you only had one bank (with room for 4 modules), it was expensive to increase the RAM, because you had to discard the old modules.

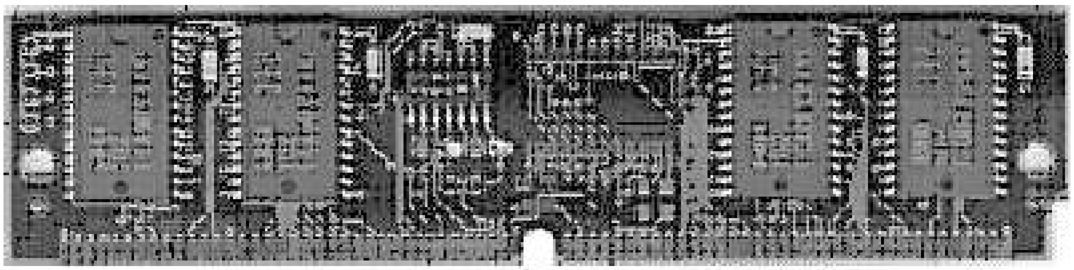
---

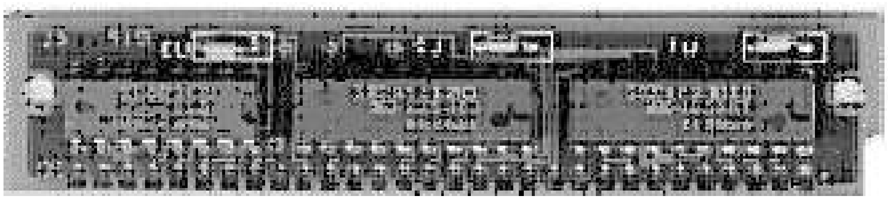
32 bit modules

With the advent of the 486 processor, demand increased for more RAM. Then the larger 32 bit modules came into use. A 486 system board could still have 4 SIMM sockets, but when the modules were 32 bit wide, they could be installed one at a time. This was quite ingenious. You could add different types of modules and still use the old ones. Also, since the 486 system board ran at only 33 MHz on the system bus, the RAM module quality was not so critical. You could mix 60 ns and 70 ns modules of different brands without problems.

Here you see a couple of SIMM modules. On top is a 64 bit module (168 pins - don't try to count them). Next is a 32 bit module with a 72 pin connector. Below is an 8 bit module with a 30 pin connector:







Number of chips per module

Some SIMM's have more chips on the module than others. Looking at just the 32-bit modules, we find modules with 2, 4, 8 or chips on each side. SIMM's with 2 MB, 8 MB and 32 MB are double sided. There are chips on both sides of the module. All these chips 16 Mbit ones.

The newest DIMM-modules holds 64 Mbit RAM chips. This way a 32 MB module is made of only 4 chips since 4 X 64 / 8 = 32.

Pentium system board with SIMM's

On the Pentium system board, the system bus is 64 bit wide. Therefore, the SIMM's are installed in pairs. Since the standard system board only has two banks with a total of four SIMM sockets, RAM expansion possibilities are limited. NOTE: never use different speed RAM modules on the Pentium system board. All modules *must* have the same speed. Here you see a few configurations on a Pentium system board with four SIMM sockets:

Bank 1	Bank 2	Total RAM

16 MB + 16 MB	-	32 MB
16 MB + 16 MB	32 MB + 32 MB	96 MB
32 MB + 32 MB	32 Mb + 32 MB	128 MB

Certain system boards (like TYAN) have 6 or 8 SIMM sockets. That provides more RAM expansion flexibility.

DIMM's

The latest RAM type, SDRAM are made in 64 bit wide modules called DIMM's. They have a 168 pin edge connector. They fit only in the newer system boards. The 82430 VX and TX chip sets can control SDRAM, as well as the LX and BX chip sets do.

Since the DIMM modules are 64 bits wide, you can install one module at a time. They are available in 8, 16, 32, 64, 128 and 256 MB, with 8, 10, and 12 ns speed. There are usually three DIMM sockets on a system board.

The advantage of SDRAM is increased speed. That allows you to increase system bus speed. With 60 ns EDO-RAM, you can run at a maximum of 75 MHZ on the system bus, while SDRAM speed can increase to at least 100 MHZ.

Some system boards have both SIMM and DIMM sockets. The idea is that you can choose between re use EDO RAM in the SIMM sockets, or install SDRAM in the DIMM sockets. They are not designed to mix RAM types although it works at some boards.

Above: a 64 MB DIMM-module holding 32 chips each of 16 Mbit (32 X 16 Mbit / 8 bit = 64 MB).

It is better to use DIMM's made of the the new 64 Mbit chips. A 64 MB module is this way made of only 8 chips (8 X 64 Mbit / 8 bit = 64 MB).

PC100 RAM

The newest DIMM-modules include a EPROM-chip holding information about the module. This chip works as a SPD (*Serial Presence Detect*) - a unit storing information about the RAM type. The idea is that BIOS can read these information and this way tune the system bus and the timings for a perfect CPU-RAM performance.

With BX chip set the system bus speed has come up to 100 MHZ. This puts new focus on the quality of the RAM modules. Hence Intel has made a new standard called PC100. Only SD-RAM modules that are constructed according to these standards are guaranteed to work at 100 MHZ. In some articles this new RAM is described at 125 MHZ SD-RAM.

Rambus RDRAM

Intel plans to use the so-called Rambus RAM (RDRAM or nDRAM) in the future. It is a advanced technology from an American company, who sells the technology to other chip manufactories for just 2% in license... And since Intel supports the RDRAM, they are going to be rich. But RDRAM should be cheap to produce, so we all profit.

Data is read in packets at a very high clock speed. 600 MHZ works fine, and GigaHertz will follow. We can drop the L2-cache if it works. The RDRAM chips have to be placed very close to the CPU to reduce radio noise.

To overview

Click & Learn. Module 3a. WWW.MKDATA.DK

About CPU's

To understand the data processing methodology, an understanding of the design and function of the CPU is essential. The following subjects will be covered on these pages. They are divided in three sub modules, which ought to be read as a unit.

You ought to read modules 2a, 2b, 2c, 2d, and 2e first.

The CPU is certainly the most important PC component. CPU stands for *Central Processing Unit*. Let us briefly study that name:

- It is a processor, because it processes data.
- It is central, because it is the center of PC data processing.
- It is a unit, because it is a chip, which contains millions of transistors.

Without the CPU, there would be no PC. Like all other hardware components, the CPU's are continually undergoing further development. You can see the explosive technological development in data processing most clearly in the development of newer and faster CPU's. The CPU's have for years doubled their performance about every 18 months, and there are no indications that this trend will stop.

When we now look at all the CPU's from a broader perspective, we can see that:

- The CPU history is closely tied to the companies IBM and especially *Intel*.
- The CPU's have their roots back to Intel's chip *4004* from 1971.
- You can identify six *CPU generations* up till today.
- The *compatibility* concept has been important throughout the development.

CPU's - brief review

CPU history starts in 1971, when a small unknown company, Intel, for the first time combined multiple transistors to form a *central processing unit* - a chip called Intel 4004. However, it was 8 years before the first PC was constructed.

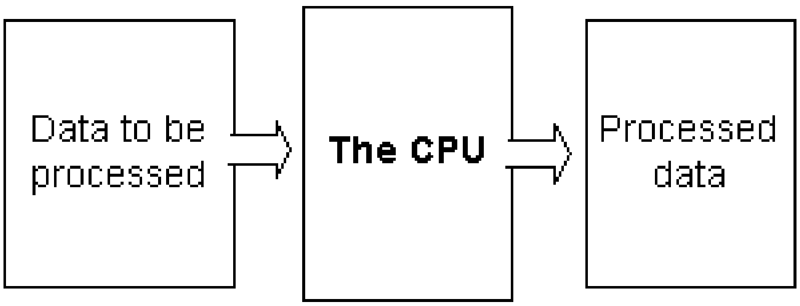
PC's are designed around different CPU generations. Intel is not the only company manufacturing CPU's, but by far the leader one. The following table shows the different CPU *generations*. They are predominantly Intel chips, but in the 5th generation we see alternatives:

PC	CPU's	Year	Number of transistors
1st. Generation	8086 and 8088	1978-81	22.000
2nd. Generation	80286	1984	128.000
3rd. Generation	80386DX and 80386SX	1987-88	250.000
4th. Generation	80486SX, 80486DX, 80486DX2 and 80486DX4	1990-92	1.200.000
5th. Generation	Pentium IBM/Cyrix 6X86 AMD K5 Pentium MMX	1993-95 1996 1996 1996-97	3.100.000 -- -- --
Improved 5th. Generation	AMD K6 Cyrix 6x86MX	1997 1997	8.800.000 5.700.000
6th. Generation	Pentium Pro Pentium II	1996 1997	5.500.000 7.500.000

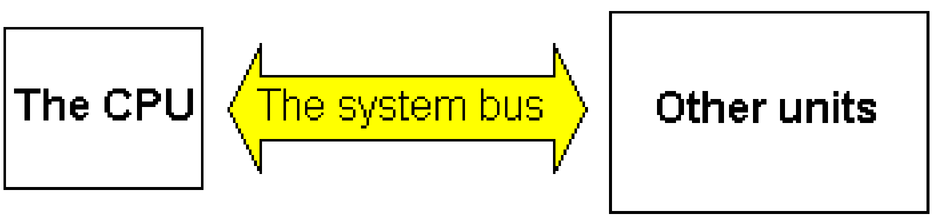
We will start by looking at what the CPU really does:

What is a CPU?

The CPU is centrally located on the system board. Since the CPU carries out a large share of the work in the computer, data pass continually through it. The CPU continually receives *instructions* to be executed. Each instruction is a data processing order. The work itself consists mostly of *calculations* and *data transport:*



Data have a path to the CPU. It is kind of a data expressway called the *system bus.* You will hear much more about this later.

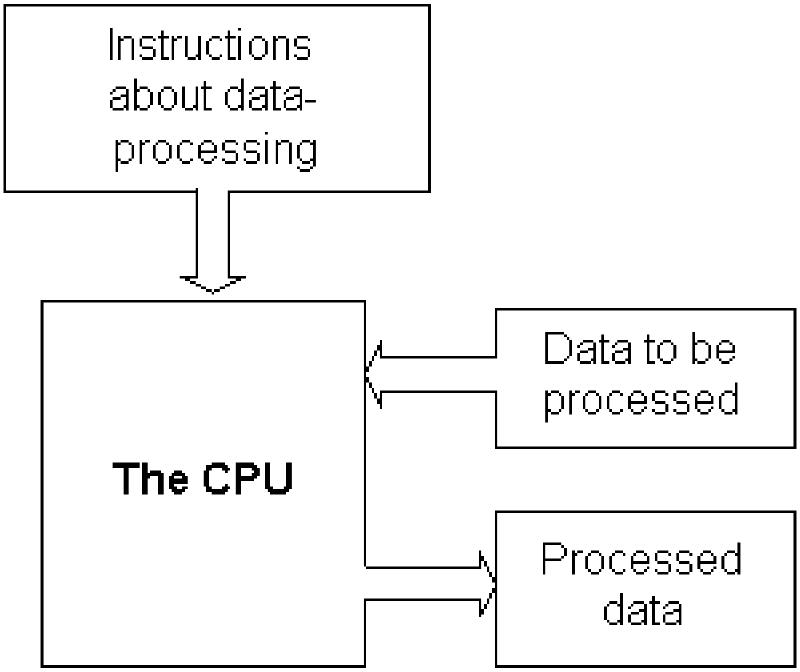


Two types of data

The CPU is fed long streams of data via the system bus. The CPU receives at least two *types* of data:

- Instructions on how to handle the other data.
- Data, which must be handled according to the instructions.

What we call instructions is *program code*. That includes those messages, which you continuously send to the PC from the mouse and keyboard. Messages to print, save, open, etc. Data are typically user data. Think about the letter, which are writing to Aunt Karen. The *contents,* letters, images, etc., are *user data*. but then you say "print," you are sending program code (instructions):



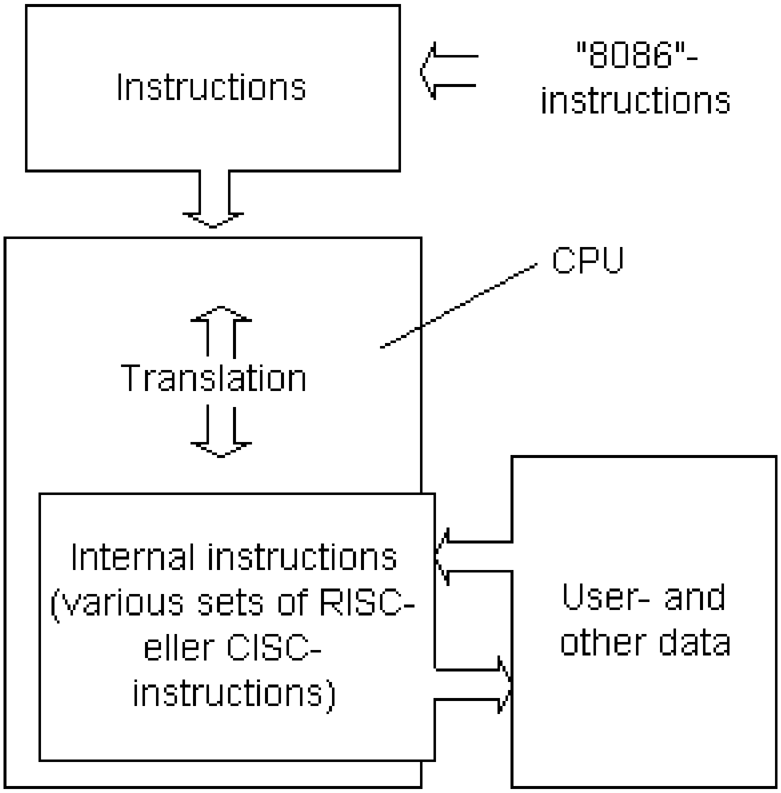
8086 compatible instructions

The biggest job for the CPU consists of *decoding* the instructions and *localizing* data. The calculations themselves are not heavy work.

The decoding consists of understanding the instructions, which the user program sends to the CPU. All PC CPU's, are "8086 compatible." This means that the programs communicate with the CPU in a specific family of instructions.

These instructions were originally written for the Intel 8086 processor, which founded the concept "the IBM compatible PC." The 8086 from 1978 received its instructions in a certain format. Since there was a desire that subsequent CPU generation should be able to handle the same instructions which the 8086 could, it was necessary to make the instruction sets compatible. The new CPU's should understand the same instructions. This *backwards compatibility* has been an industry standard ever since. All new processors, regardless of how advanced, must be able to handle the 8086 instruction format.

Thus, the new CPU's must use much effort to translate the 8086 instruction format to internal instruction codes:



CISC and RISC instructions and their handling

The first CPU's had a so called *Complex Instruction Set Computer* (CISC). This means that the computer can understand many and complex instructions. The X86 instruction set, with its varying length from 8 to 120 bit, was originally developed for the 8086 with its mere 29000 transistors.

Reduced Instruction Set Computer (RISC): The RISC instructions are brief and the same length (for example 32 bit long, as in Pentium Pro), and they process much faster than CISC instructions. Therefore, RISC is used in all newer CPU's. However, the problem is that the instructions arrive at the CPU in 8086 format. Thus, they must be decoded

For every new CPU generation, the instruction set has been expanded. The 386 came with 26 new instructions, the 486 with 6 new instructions, and Pentium with 8 new instructions. These changes mean that some programs require at least a 386 or a Pentium processor to work.

There is also a continuous optimizing of the instruction handling process. One is that the clock frequency increases, as we will see later - the faster, the better. But what can the CPU do in one clock tick. That is critical to its performance. For example, a 386 needed 6 clock ticks to add a number to a sub total. A job which the 486 manages in only two clock ticks, because of more effective instruction decoding, 5th and 6th generation CPU's can execute more than one of those operations in one clock tick, since they contain more processing lines (pipelines), which work parallel.

Read also the section about MMX instructions.

Floating-point unit - FPU

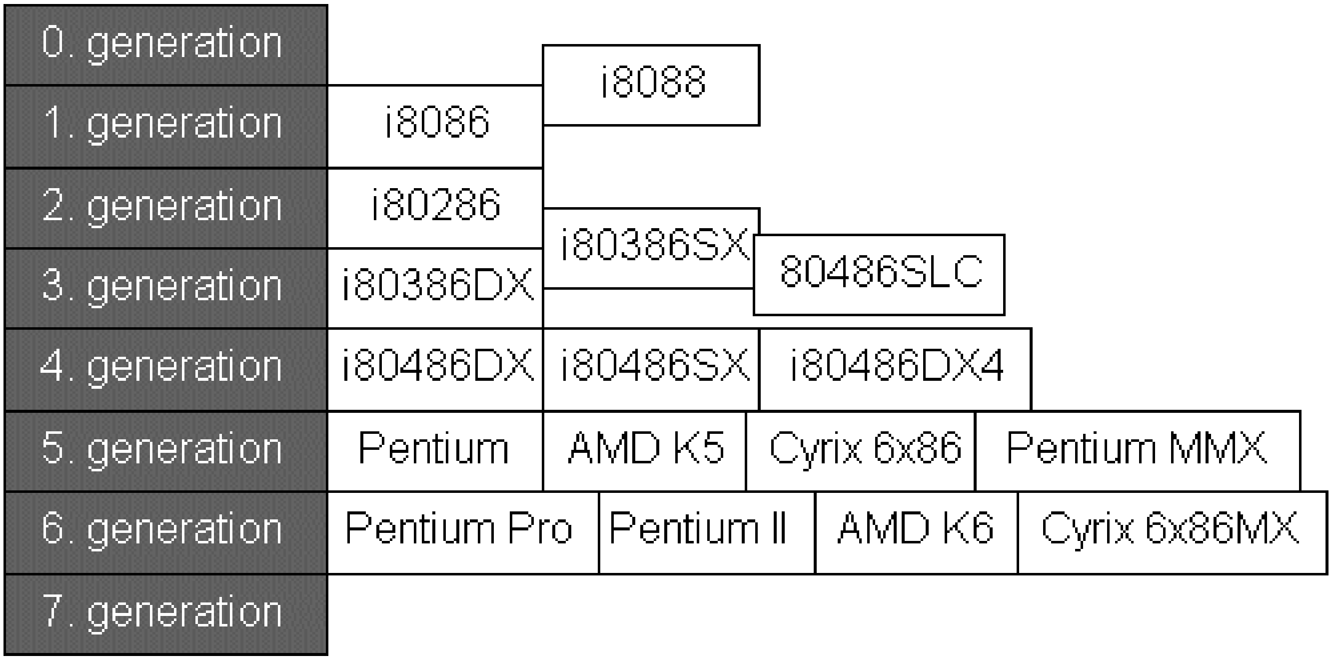
The first CPU's could only work with whole numbers. Therefore, it was necessary to add a mathematical co-processor (FPU), when better math power was needed. Later, this FPU was built into the CPU:

CPU	FPU
8086	8087
80286	80287
80386	80387
80486DX	Built in
80486SX	None
Pentium and thereafter	Built in

It is said that Intel's CPU's have by far the best FPU units. Processors from AMD and Cyrix definitely have a reputation for providing sub standard performance in this area. But, you may not utilize the FPU. That depends on the applications (user programs) you are using. Common office programs do not use the floating point operations, which the FPU can handle. However, 3D graphics programs like AutoCad do. Therefore, if you use your PC in advanced design applications, the FPU performance becomes significant. For most users, it is only of limited importance.

Many brand names

As mentioned, there are CPUs of many brand names (IBM, Texas, Cyrix, AMD), and often they make models, which overlap two generations. This can make it difficult to keep of track of CPU's. Here is an attempt to identify the various CPU's according to generation:



Now, let us see how CPU speed has been improved through generations. Click for Module 3b.

To overview

Click & Learn. Module 3b. WWW.MKDATA.DK

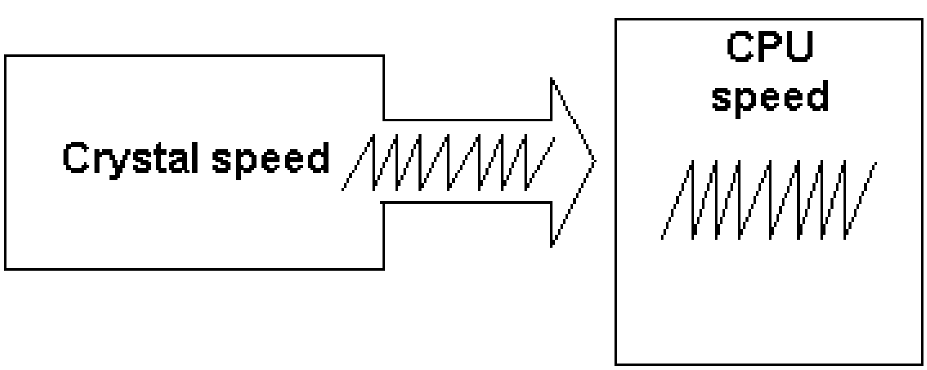
The CPU – developments and improvements

If you have to improve a CPU – and that happens all the time – it is not only a matter of technical development. There are many bottlenecks in and around the CPU, which are continually being bettered.

To understand these technological improvements, one must remember that the CPU is a data processing gadget, mounted on a printed circuit board (the system board). Much of the data processing takes place inside the CPU. However, all data must be transported to and from the CPU via the system bus. But what determines the speed of the CPU?

Clock frequency

We know this from the ads: "A Pentium 166 MHZ." The 166 MHZ is the *clock frequency.* Actually, there is a small crystal on the system board. which continually ticks to the CPU at a steady number of *clock ticks* per second. At each clock tick *something happens* in the CPU. Thus, the more ticks per second – the more data are processed per second.



The first CPU's worked at a frequency of 4,77 MHZ. Subsequently then, clock frequencies rates rose to 16, 25, 50, 66, 90, 133 and 200 MHZ to the best today, which probably operate at 266 MHZ. Clock frequencies are still being increased. In a few years we will have CPU's operating at

400 and 500 MHZ.

To reach these very high clock frequencies, one has to employ a technique called clock doubling :

Clock doubling in the CPU

The problem with the high clock frequencies is to ensure that other electronic components keep up with the pace. It is rather simple to make data move very fast inside a chip where the print tracks are microscopic. But when we move outside the chip, other problems appear. The other components must be able to keep up with the pace. When the frequency gets too high, the circuit board print tracks start acting as antennae and various forms of "radio noise" appears. Briefly, it becomes expensive to make the rest of the hardware to keep up with these high frequencies.

The solution to this problem was to split the clock frequency in two:

- A high internal clock frequency, which governs the pace of the CPU.
- A lower external clock frequency, which governs the pace on the system bus. This is where the CPU exchanges data with RAM and the I/O units.

Intel's 80486DX2 25/50 MHZ was the first chip with clock doubling. It was introduced in 1992 with great potential. For a lower price you could acquire a chip, which provided 90% of the 486DX50 performance. The DX50 runs at 50 MHZ both internally and externally. The DX2 runs at just 25 MHZ on the system bus. This enables lower cost system boards. Also RAM speed demands are much lower.

Clock doubling occurs inside the CPU. If the system board crystal works at 25 MHZ, the CPU will receive a signal every 40 nanosecond (ns). Internally in the CPU, this frequency is doubled to 50 MHZ. Now the clock ticks every 20 ns inside the CPU. This frequency governs all internal transactions, including *integer unit*, *floating point unit*, and all *memory management unit* operations as well as others. The only area still working at 25 MHZ are external data transfers. That is transfers to RAM, BIOS and the I/O ports.

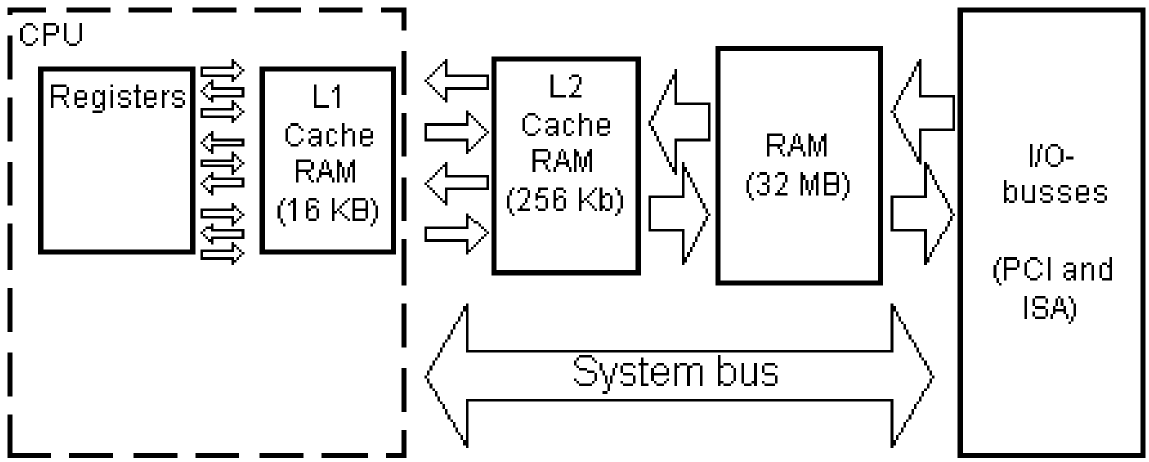
Today the speed problem is in RAM. The ordinary FPM RAM and EDO RAM can function at a maximum of 66 MHZ (possibly 75 MHZ). Therefore, Pentium and similar CPU's are "clocked up" 2-4 times internally. They work well at high frequencies like 166, 200, 233 and 266 MHZ.

Please read Module 3d for more about clocking.

About CPU cache RAM

The CPU must deliver its data at a very high speed. The regular RAM can not keep up with that speed. Therefore, a special RAM type called *cache* is used as a buffer - temporary storage. To get top performance from the CPU, the number of outgoing transactions must be minimized. The more data transmissions, which can be contained inside the CPU, the better the performance. Therefore, the 486 was equipped with a built in mathematical co-processor, *floating point unit* and 8 KB L1-cache RAM. These two features help minimize the data flow in and out of the CPU.

Cache RAM becomes especially important in clock doubled CPU's, where internal clock frequency is much higher than external. Then the cache RAM enhances the "horsepower" of the CPU, by allowing faster receipt or delivery of data. Beginning with 486 processors, *two* layers of cache are employed. The fastest cache RAM is inside the CPU. It is called L1 cache. The next layer is the L2 cache, which are small SRAM chips on the system board. See at the illustration below:



Cache overview

L1-cache first appeared in Intel's 80486DX chip:

CPU	Cache size in the CPU
80486DX and DX2	**8 KB L1**
80486DX4	**16 KB L1**
Pentium	**16 KB L1**
Pentium Pro	**16 KB L1 + 256 KB L2**
Pentium MMX	**32 KB L1**
AMD K6	**64 KB L1**
Pentium II	**32 KB L1**

Today, bigger and better CPU cache is a natural step in the development of new CPU's.

Areas of development

In the following table, you see some of the technologies, which can be improved in the CPU design. Note that internal means inside the CPU. External speed, etc. refers to features immediately outside the CPU – on the system board.

Development area	Significance	Example
Internal clock frequency	**speed of data processing *inside* the CPU.**	**266 MHZ**
external clock frequency	**Speed of data transfer to and from the CPU via the system bus.**	**66 MHZ**
Clock doubling	**That the CPU works x times faster internally than externally.**	**4,0 times (like above)**
Internal data width	**How many data bits can the CPU process simultaneously.**	**32 bits**
External data with	**How many data bits can the CPU receive simultaneously for processing**	**64 bits**
Internal cache (Level 1 cache)	**Large and better L1-cache, which is a small fast RAM. It works as a buffer between CPU and regular RAM.**	**32 KB**
Instruction set	**Can the instruction set be *simplified*, to speed up program processing? Or can it be improved?**	**RISC code** **More pipelines** **MMX instructions**

The CPU – speed measurement

When we look at an individual CPU, its speed is the most significant feature. All newer CPU's can do the same. You can run Office 97 in Windows 95 on a 386 CPU. It would be terribly slow, but it *is* possible.

Speed is the primary difference between newer CPU's. Speed improvement is a product of the above mentioned technologies (such as clock frequency and bus width).

There are many, many ways to measure CPU speed. The subject is boundless. For years, Norton's Speed Index has provided a good sounding board. That is a test, which can be run on any PC with the Norton Utilities Sysinfo program.

In the table below, you see a number of the most common older CPU's. You can see how they are designed regarding clock speed and bus width. The last column shows their *Norton Speed Index* (SI). That is a relative number, which can be used to compare different CPU's.

CPU	CPU speed	Clock doubling	System bus speed	Data width	SI
8086	4,77 MHZ	1	4,77 MHZ	16 bit	1
80286	12 MHZ	1	12 MHZ	16 bit	8
80386DX	25 MHZ	1	25 MHZ	32 bit	40
486 DX2-66	66 MHZ	2	33 MHZ	32 bit	142
486 DX4-133	133 MHZ	4	33 MHZ	32 bit	288
Pentium 75	75 MHZ	1,5	50 MHZ	64 bit	235
Pentium 90	90 MHZ	1,5	60 MHZ	64 bit	278
Pentium 100	100 MHZ	1,5	66 MHZ	64 bit	305
Pentium 133	133 MHZ	2	66 MHZ	64 bit	420
Pentium 166	166 MHZ	2,5	66 MHZ	64 bit	527
Pentium 200	200 MHZ	3	66 MHZ	64 bit	629

Read also about the system bus in module 2b.

CPU changes - historical review

This describes briefly the changes throughout the CPU generations:

8088 and 8086: The first PC CPU was the 8088. It is a 16 bit CPU, but only internally. The data bus width is only 8 bit, to facilitate the manufacture of low cost PC's. Actually the 8088 is a 16/8 bit CPU. Logically it could have been named 8086SX. The 8086 was the first total 16 bit CPU in this family.

8086 to 80286: The 286 was a big advance, relative to the first generation chips. Clock frequency was increased, but the major improvement was in optimizing instruction handling. The 286 produced much more per clock tick than 8088/8086 did. Another innovation was the ability to run in *protected mode* - a new work mode, which pointed towards the shift from DOS to Windows.

80286 to 80386: The change to the 386's came in the late 80's. 80386 was the first 32 bit CPU. From the traditional DOS PC's point of view, this was not a revolution. A good 286 ran as fast as the first 386SX's - despite the implementation of 32 bit mode.

80386SX: A discount edition of 386DX. It has only 16 bit external data bus contrary to the DX 32 bit. Also, the SX has only 24 address lines, Therefore, it can only address a maximum of 16 Mb RAM. It is not really a true 386.

80386 to 80486: Generally speaking, the 486 runs twice as fast as its predecessor - all things being equal. That is because of better implementation of the x86 instructions. They are handled faster, more in RISC mode. At the same time bus speed is increased, but both 386DX and 486DX are 32 bit chips. A novelty in the 486 is the built in math co-processor. Before, that had to be installed as a separate 387 chip**.**

80486SX: This was a new discount chip. The math co-processor was simply omitted.

Cyrix 486SLC: Cyrix and Texas Instruments have made a series of 486SLC chips. They used the same set of instructions as did the 486DX, and they run at 32 bit internally, like the DX. However, externally they run at only 16 bit (like a 386SX). Therefore, they can only handle 16 MB RAM. Furthermore, they only have 1 KB internal cache and no mathematical co-processor. Actually they are just improved 286/386SX's. They are not cloned chips. There are substantial differences in their architecture compared to the Intel chips.

IBM 486SLC2: IBM had their own 486 chip production. The series was named SLC2 and SLC3. The latter was also known as *Blue Lightning*. These chips could be compared to Intel's 486SX, since they did not have a built-in mathematical co-processor. However, they had 16 KB internal cache (compared to Intel's 8). What reduced their performance was the bus interface, which was from the 386 chip. SLC2 runs at 25/50 MHZ externally and internally, while the SLC3 chip runs at 25/75 and 33/100 MHZ. IBM manufactured these chips in their own facilities, licensed by Intel. IBM may change the architecture, but these chips may only be used in IBM PC's and system boards. They may not be sold separately.

Further 486 developments

DX4: Intel's DX4 processors represent an improvement on the 80486 series. Their speed is tripled from 25 to 75 MHZ and from 33 to 100 MHZ. Another DX4 chip is speeded up from 25 to 83 MHZ.

Contrary to what you might think, the DX4 are not named for a quadrupling. They are so named because of the registry of Intel's 80486 and 80586 names. The DX4 name is separated from that context, so it could be patented. If DX3 referred to a tripling, this would not work. The same type of problem caused the next generation chip to be named Pentium, rather than 80586.

The DX4 has 16 KB internal cache and operates on 3.3 volt (they will tolerate 5 volt, to accommodate existing system boards). DX and DX2 have only 8 KB cache and require 5 volt with inherent heat problems.

5X86: AMD has made a series of so called 5X86 CPU's. Those are improved 486's, which approach the 5th generation chips, hence their name. Their 120 MHZ model is noteworthy. It can easily be tuned to run at 160 MHZ.

Continue the description by looking at 5th and 6th generation chips, Pentium etc.: Click for Module 3c

---

To overview

Click & Learn. Module 3c. WWW.MKDATA.DK.

About 5th and 6th generations CPU's

With Intel's Pentium from 1993, a new era began in the continued CPU development. In these pages, we will look at different variations and further development of 5th. generation CPU's.

Pentium Classic (P54C)

This chip was developed by Intel in Haifa, Israel. The processor is super scalar, meaning that it can execute more than one instruction per clock tick. Typically, it handles two instructions per tick. In this respect, we can compare it to a double 486. At the same time there have been big changes in the system bus: the width is doubled to 64 bit and the speed is increased to 60 or 66 MHZ. This results in a substantial improvement from the 486 technology.

Originally, Pentium came in two versions: a 60 MHZ and a 66 MHZ. Both operated on 5 Volt. This produced a lot of heat (it was said that you could fry an egg on them!) The next Pentium (P54C) generation had a built in 1½ clock doubling, and ran at 3½ Volt. This took care of the heat problem.

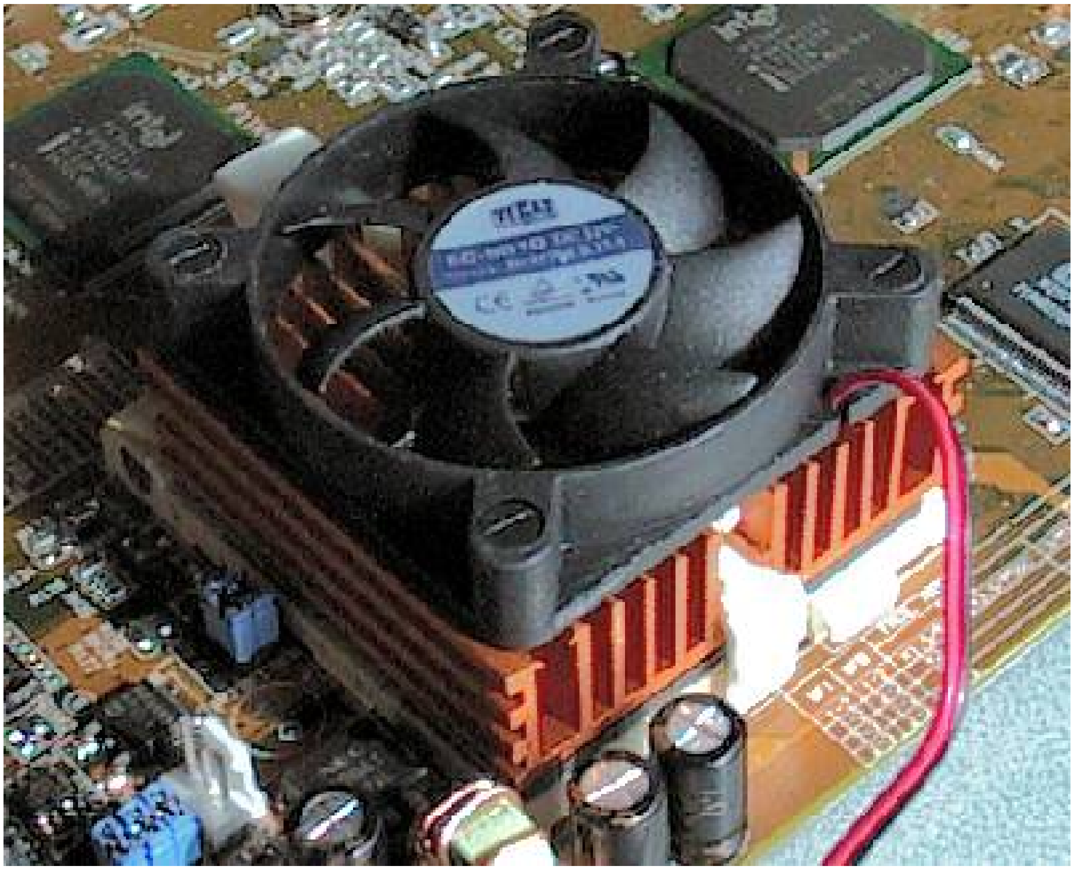
Since then, Intel carried two Pentium lines: those which run at 60 MHZ on the system bus (P90, P120, P150, and P180) and the best, which run at 66 MHZ (P100, P133, P166 and P200).

Cooling

All modern CPU's share a common need for cooling. Make sure to include a good cooler. It has to be matched to the size of the CPU.

- It has to be attached properly, either with glue or a clamp, which fits the CPU.
- It must have a substantial size heat sink - the bigger the better.
- The fan must be mounted in roller bearings, to minimize noise.

The bigger the fan and heat sink, the better it is. The CPU will operate more reliably. It will have a longer life span, and it can possibly be over clocked. If you buy Intel CPU's, buy them "in a box". It is a special package, priced slightly higher than just the CPU. They always include a good fan and a three year warranty.



Pentium with fan. Photo taken with Canon Powershot 600. JPG-file 1:30, 32 KB.

Pentium MMX (P55C)

MMX is a new set of instructions (57 new integer instructions, four new data types, and eight 64 bit registers), which expand the capabilities of the CPU. It is an addition to the original Pentium set of instructions.

The new instructions are designed for multimedia programs. The programmers can utilize these instructions in their programs. These allow the Pentium to provide improved program execution. MMX is a new standard, which Intel will include in all their CPU's. Both Cyrix and AMD use MMX in their 6th generation CPU's (K6 and M2). Programs, which are written with MMX instructions, can still be run on, for example, a Pentium without MMX. However, execution is slower with the traditional instructions.

More L1 cache and higher clock frequency

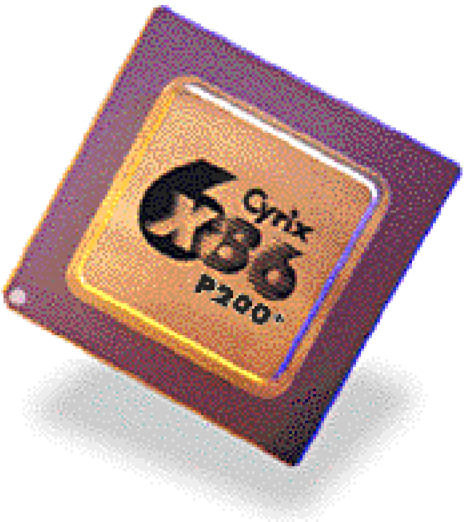
The new P55C Pentiums are further improved with 32 KB L1 cache (the old had 16 KB). This in itself is a good feature. There are also other improvements in the CPU. These improvements should provide 10-20% better performance than the old CPU's.

Dual voltage

P55C requires a new system board. Not because of MMX - that is pure software, but because of changes in the power supply. P55C operates with *dual voltage* technology. To reduce heat generation, this chip requires two different voltages: 2.8 Volt to *the nucleus* and 3.3 Volt to the I/O section. The old system boards for the P54C's have only one voltage to the CPU. Thus, the new CPU requires a new system board.

Cyrix 6x86

This is a low cost alternative to Pentium. The chip from the Cyrix company, which was purchased by IBM in 1996, is a cheap Pentium copy.



The chip is Pentium compatible, since it fits in a socket 7, which is the Pentium connection to the system board. When Cyrix suggests a 6th generation, it is because the 6X86 employs advanced techniques, which are not found in Intel's Pentium. Thereby Cyrix gets improved performance from their chip with the same clock speed. They market their CPU's with a comparison to Intel's clock frequency.

As mentioned, the 6x86 chips have a lower internal speed than their name suggests. Below, you can see the data for the different models:

Cyrix model	CPU speed	Clock doubling	System bus speed
P120+	100 MHZ	2	50 MHZ
P133+	110 MHZ	2	55 MHZ
P166+	133 MHZ	2	66 MHZ

P200+	**150 MHZ**	**2**	**75 MHZ**

Cyrix 6X86 are known for poor performance regarding floating point operations. Therefore, If you work with AutoCAD or other 3D programs, 6X86 would hardly be your choice. There have also been problems with Cyrix and NT 4.0. In my experience, they work well with common office programs in Windows 95. I was very satisfied with the P166+ I had. Of course I would prefer a genuine Pentium 166, but I was not willing to pay three times the price at that time.

The 6X86 has since been improved with Dual Voltage (like Pentium P55C). This reduces power consumption and heat generation.

AMD

AMD is another CPU brand. Their Pentium-like chips offer Intel tight competition. AMD use their own technologies Therefore, they are not clones. They have these series:

- K5, corresponding to the earlier Pentiums without MMX
- K6, which compete with Pentium MMX and Pentium II
- K7, which comes in 1999 and is not socket7 compatible.

K5

Their K5 is found for example as PR133. It performs almost like a Pentium P133, but runs at only 100 MHZ internally. It still has to be installed in the system board like a P133.

AMD's K5 also exists as PR166. As the name suggests, it is intended to compete with Intel's P166. It is interesting that it runs at only 116.6 MHZ internally (1.75 X 66 MHZ). According to the highly respected German magazine *c't,* issue *3.97 page 20,* it actually runs at least as fast as the P166. This is due to an optimized cache and other new developments. The only feature on which it cannot match the P166 is in floating point operations. These are typically necessary in 3D calculations in AutoCAD and similar applications. PR133 and PR166 cost far less than the similar Pentium models.

K6

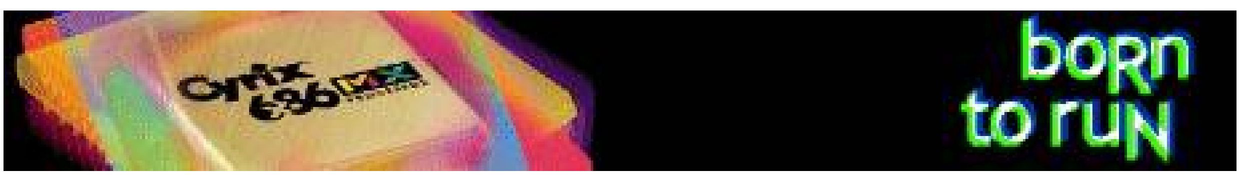
AMD's K6 is from 1997. This chip performs better than Pentium MMX . It is equipped with a 32+32 KB L1 cache and MMX. It contains 8.8 million transistors!

It is exciting that the K6 (like K5) is prong compatible with Pentium. Thus, it can be mounted in a socket 7 on a regular Pentium system board. On the older system boards, it is possible that the BIOS has to be updated to make it work. However, it is a positive trend that new CPU's are becoming available for use in the well-tested and inexpensive socket7 system boards. K6 performs best when the BIOS recognizes the chip. Then its full potential can be utilized. That requires the newer type dual voltage system boards. The K6-200 requires 2.9 volt for its core. The other models require 2.8 volt as the Pentium MMX.

AMD has actually created a hybrid chip, which is placed between 5th and 6th generation CPU's. Intel will concentrate exclusively the development of their 6th generation chip Pentium II. So AMD has announced that they will develop new chip sets for the socket 7 system boards.

So far they produce K6 model 7 "Little Foot" running at 266 and 300 MHZ. These high performance K6's are sold at very reasonable prices. Their problem seems to be to produce enough chips. These chips run on just 2,2 Voltage, they hardly need a cooler....

Cyrix 6X86MX (M2)



Cyrix also has a high performance chip, placed between 5th and 6th generation. It was announced as M2, but is now called 6X86MX, It is also prong compatible with Pentium. This gives additional possibilities to assemble PC's on ordinary socket 7 system boards.

6X86MX has 64 KB internal L1 cache, which is very impressive. They also utilize technologies which are not found in Pentium MMX. These chips are so named, to compare them with genuine Pentiums, although their internal clock speed is lower than corresponding Intel processors:

6X86MX	Internal speed	External speed
PR166	150 MHZ	60 MHZ
PR200	166 MHZ	66 MHZ

PR233	**188 MHZ**	**75 MHZ**
PR266	**? MHZ**	**? MHZ**
PR300	**? MHZ**	**100 MHZ ?**

6x86MX is unique compared to the other 6. generation CPUs (Pentium II and Pro and K6) since it does not work upon a RISC kernel. 6x86MX executes the original CISC instructions as does the Pentium MMX.

The 6x86MX has plenty of internal registers:

CPU	Number of 32 bit CPU registers
Pentium MMX	8
6x86MX	32
Pentium Pro	40
K6	48

The 6x86MX has - as all processors from Cyrix - a problem concerning the FPU unit. However, using standard office applications, this is of no concern.

At this time, the 6x86MX is a powerful CPU that offers more value for money than Intel's Pentiums do! It is evident that Cyrix intends to continue this line of processors, and this definitely is a positive trend. Intel gets competition, and it keeps the well-tested and inexpensive socket7 system boards in the market.

Two brands of 6x86MX

The 6x86MX processor is produced by National/Cyrix as well as by IBM. The architecture is the same, but the chips are built at different plants. Their top-models are (Jun '98):

CPU	Cyrix PR300	IBM PR333
Technology	0.35 micron	Semi 0.25 micron
Clock speed	3.5 X 66 MHZ = 233 MHZ	3.0 X 83,3 MHZ = 250 MHZ

IBM uses a completely new technology for the PR333 chip. It is patented and called Flip-Chip. The die is soldered directly to the ceramic casing and this causes less induct ions. IBM is preparing for real 0.25 micron processing technology later this year, which will increase the external clock speed to 100 MHZ.

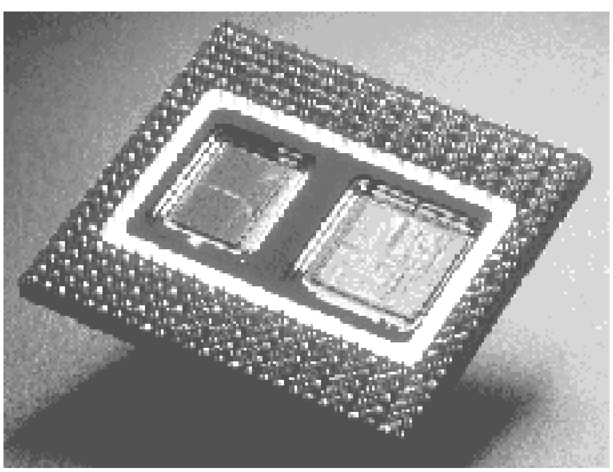
Pentium Pro

Pentium Pro is a pure RISC processor. It is optimized for 32 bit processing in Windows NT or OS/2. It is unique in that the L2 cache is built-in. This is like two chips in one.

Pentium Pro development started in 1991, in Oregon. It was introduced in late 1995. The new features were:

- Built in optimized L2 cache with 256 KB or 512 KB. This is connected to the CPU itself with a 64 bit back side bus. Thus, the L2 cache runs synchronous with the CPU speed.
- Multiple branch prediction, where the CPU anticipates the next instruction. Data Flow Analysis, which should reduce data dependence. Speculative Execution, where the CPU attempts to anticipate instruction results.
- 5.5 million transistors in the CPU, 15 million for the 256 KB SRAM L2 cache. (6 transistors per bit).
- 4 pipelines for simultaneous instruction execution.
- RISC instructions with concurrent x86 CISC code to MicroOps RISC instructions decoding.
- 2.9 Volt 4 layer BiCMOS processor technology.
- Patented protocol. Thus, other CPU manufacturers cannot use the Pentium Pro socket and chip set. This is not to the users advantage.

Here you see a rectangular chip. The CPU and L2 cache are separate units inside this chip:



Pentium Pro is primarily optimized to 32 bit program execution. Since I use a PPro 200 MHZ (which runs at 233 MHZ) and experience tremendous power in my Windows 95 environment, I can recommend it for ordinary use.

Pentium Pro is especially designed for Windows NT and other genuine 32 bit environments.

Since the introduction of Pentium II, the price on PPro has dropped and soon it will be out of production.

Pentium II

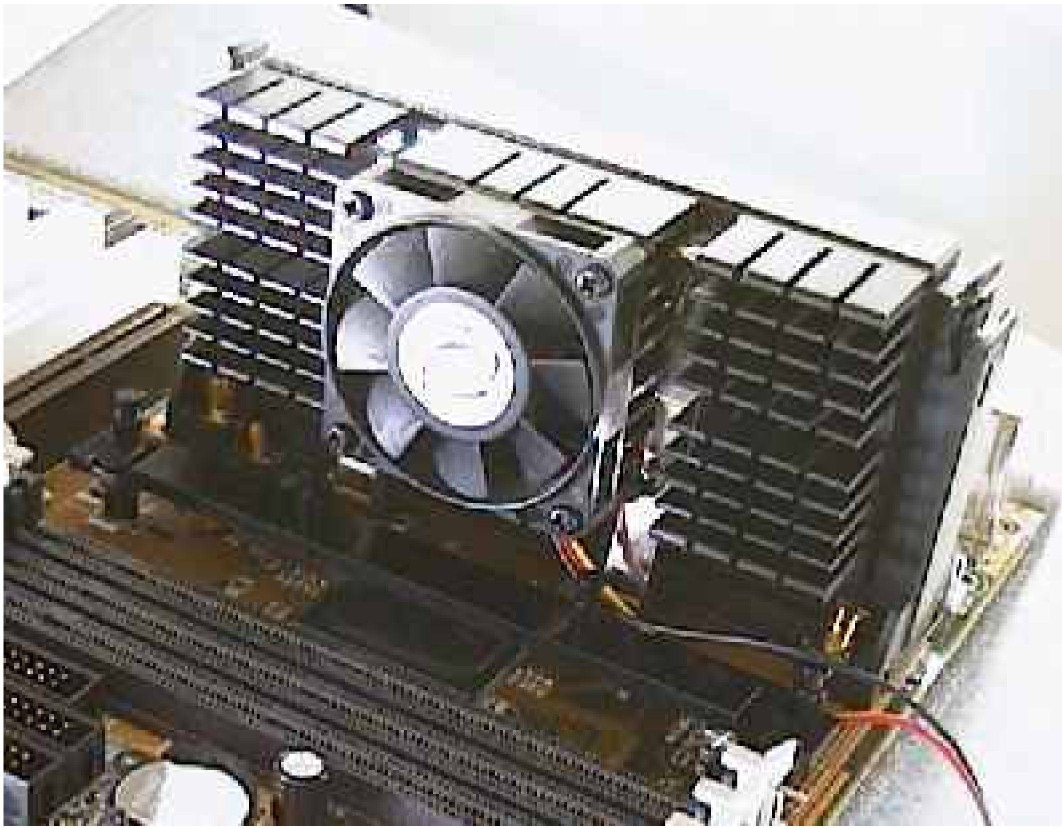
Pentium Pro "Klamath" was the code name for Intel's new top processor. It is a partially reduced and partially improved Pentium Pro model. The construction of Pentium II is interesting but also controversial.

- **With MMX instructions**
- **Improved 16 bit program execution (joy for Windows 3.11 users)**
- **Doubled and improved L1 cache (16 KB + 16 KB)**
- **CPU is mounted together with 512 KB L2 on a cartridge connected to the main board using the *slot one* connector.**
- **New increased internal speed: from 233 MHZ to 333 MHZ**

The most interesting change is the separation of CPU and L2 cache. Intel found it too costly to combine them in one chip as in Pentium Pro. To facilitate mass production, cache RAM of a different brand (Toshiba) was used. The cache RAM is marked 7 ns allowing a clockfrequency of maximum 150 MHZ.

Pentium II is a large rectangular plastic box, which contains the CPU and cache. There is also a small controller (S82459AB) and a well dimensioned cooling fan. All are mounted on a card. This card with chips weighs about 380 g (13 ounces). It fits in a new 242 pin Single Edge Connector on the system board.

Here you see the SEC-module:



With the new design, the L2 cache has its own bus. It runs at half the CPU speed, like 133 MHZ. This is clearly a retrogression from the Pentium Pro, which can run at 200 MHZ between the CPU and L2 cache. It is countered by the improved L1 cache, which really zips along! Here you see a comparison:

CPU	L1 transfer rate	L2 clock speed	L2 transfer rate
Pentium 200	**777 MB/sec.**	**66 MHZ**	**67 MB/sec.**
Pentium 200 MMX	**790 MB/sec.**	**66 MHZ**	**74 MB/sec.**
Pentium Pro 200	**957 MB/sec.**	**200 MHZ**	**316 MB/sec,**
Pentium II 266 MHZ	**1.175 MB/sec.**	**133 MHZ**	**221 MB/sec.**

Pentium II is Intel's top model. It is currently available in 233, 266, 300, 333, 350 and 400 MHZ editions. With the new 82440LX and 82440BX chip sets Pentium II is an excellent performer. Read on for more information. But first the rival:

AMD K6-II or K6-3D

The newest version of the K6 code name "Chomper" is an intriguing new CPU. The K6 model 8 is called K6-2, and it is manufactured with 0.25 micron technology. This gives the front side bus (system bus) a speed of 100 MHz.

The K6-2 also holds a new 3D plug-in (called 3DNow!) for better game performance. It is 21 new instructions that can be used by software developers giving a better 3D-performance. To benefit from it, you need a graphics driver ore a came, which deals directly with the new commands. 3DNow! is not compatible with MMX.

The next version of this chip - model 9 code name "Sharptooth" - will incorporate three levels of cache! In the CPU you find both a L1- and a L2-cache. The in-chip L2 cache of 64 KB runs at processor speed. On the main board you will find L3-cache up to 1 megabyte!

Deschutes

On January the 26th Intel introduced the new 333 MHZ model of Pentium II. It is the first of a second generation Pentium-II's known under the code name "Deschutes". The chips are produced with 0.25 micron technology, which reduces the power consumption with more than 50 % compared to the original Pentium II "Klamath" with its 0.35 micron technology. The core voltage is down from 2.8 to 2.0 Volt.

On April the 15th Intel released the next line of Deschutes. The system bus is sped up to 100 MHz. This will internally be multiplied by the clock factors 3.5, 4.0 and (June 1998) 4.5, making the CPU running at 350, 400 and 450 MHz. These CPU's uses the new chip set: 82440BX.

So the Deschutes chips use two different mother boards:

- LX-based for the 333 MHZ version (5 X 66 MHZ)
- BX-based for the 350 and 400 MHZ versions (3.5 and 4.0 X 100 MHZ).

The L2 cache RAM has to deal with these higher clock frequencies:

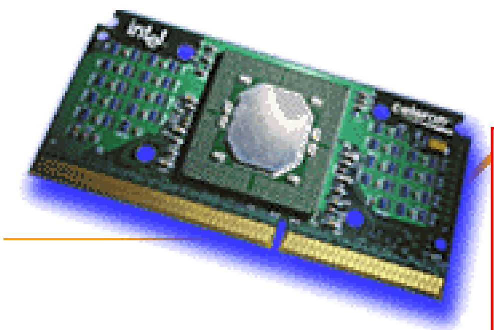
CPU Clock	RAM type	Controller
333 and 350 MHz	5.5 ns	82459 AC tag RAM
400 MHz	5.0 ns	82459 AD tag RAM

Also the cache RAM chips are *cooled* on the new cartridges.

The Celeron, a Pentium II-SX

Intel is having a hard time with the Pentium II which is to expensive. Most users buy the AMD K6-233, which offers very good performance at a moderate price.

Thus Intel has created a brand new CPU called Celeron. It is a Pentium II cartridge except from the L2-cache, which has been chopped away. It uses a new Covington core, and we could just as well call it a Pentium II-SX. Later this year Intel replaces their Pentium MMX with the Celerons.



This inexpensive Celeron cartridge fits into Slot 1 and it runs at 66 MHZ FSB. The internal clocks are at 266 or 300 MHZ and delivers very good performance for games, especially. Concerning office applications, the lack of L2-cache is a great disadvantage.

The Celeron probably will be fine for overclocking, since many of the problems here arises from the onboard L2 cache. The L2 cache RAM cannot function at high clock frequencies, but without L2 cache RAM this problem does not occur with the Celeron.

The later 333 Mhz version of Celeron holds 128 KB L2 cache inside the CPU it self, using the new Mendocino core.

Pentium II Xeon

On July 26th 1998 Intel will introduce a new Pentium II edition. The processor is named Pentium II Xeon.

It will require a new version of Slot One (*Slot Two*). The module will have the same dimensions as the current Pentium II, but there are a few important innovations and improvements:

- The CPU will mounted in a new Slot Two with three layers of edge connectors.
- 100 MHZ system bus with clock doubling 4.0 and 4.5 (meaning 400/450 MHZ clock speed internally).
- New type L2 cache: CSRAM, which can run at full CPU speed.
- 512, 1,024, or 2,048 KB L2 RAM.
- Up to 4 GB RAM can be in cache.

The Xeon chip is for high performance servers. The first topmodel will hold 2 MB L2 cache on the cartridge, running at full 450 MHZ. This chip will cost $4,500!

Katmai and further

In 1999 Intel will come up with a new enhanced MMX2 set of graphics instructions. This will be called Katmai and is intended to speed up 3D gaming performance - just as AMD's 3DNow! technology. Katmai includes "double precision floating point single instruction multiple data" (or DPFS SIMD for short).

Later in 1999 follows a chip with the code name Tanner or Willamette, which also is 32 bit Pentium II-styled CPU.

Merced

Code name for a 7th. generation CPU, which Intel is developing together with the HP company, HP has great experience in the the manufacture of high end CPU's (RISC). I will tell you about that, as the information becomes available.

- 64 bit CPU, will require a new 64 bit operating system (Windows 2000?).
- Clock frequency: 1000 MHz (1 GHZ).
- "Massive hardware units": 128 integer and 128 floating point registers with multiple integer and floating point units all working parallel.
- 0.18 micron technology.
- Slot M catridge.

After Merced follows a CPU with code name McKinley. And so on...........

CPU sockets and chip sets

There are currently three different CPU sockets. Here you see them described with the latest chip sets:

Socket	Fits CPU	Chip set
Socket 7	Pentium, MMX, K5, 6x86, K6, 6x86MX	82430TX
Socket 8	Pentium Pro	82440FX
Slot One	Pentium II	82440FX 82440LX
Slot One	Pentium II (100 MHZ system bus)	82440BX
Slot Two	Pentium II Xeon	82450NX

Only Socket7 may be copied freely. The other two are Intel's patents. They may be manufactured by others on license from Intel.

Voltages - dual voltage

One of the newest CPU technologies is the continually thinner wires inside the chip. With thinner wires, the CPU can operate at lower voltage. That results in a smaller CPU generating less heat and with the ability to operate at higher speeds. A step in this development is the design of dual voltage chips:

- The interface to the I/O bus, which always requires 3.3 volt.
- In internal CPU parts, it is advantageous to reduce the voltage as much as possible. This can be done because of the extremely thin wires in the CPU.

The new socket 7 system boards have a two part voltage regulator to match the needs of the CPU. Here are some selected CPU's and their voltage requirements:

CPU	**Internal voltage**	**I/O voltage**
Pentium MMX	**2.8 Volt**	**3.3 Volt**
AMD K6	**2.8/2.9 Volt**	**3.3 Volt**
Cyrix 6X86MX	**2.8 Volt**	**3.3 Volt**
Pentium II "Klamath"	**2.8 Volt**	**3.3 Volt**

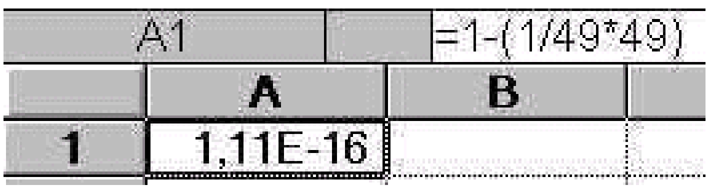
AMD K6-266 +	**2.2 Volt**	**3.3 Volt**
Pentium II "Deschutes"	**2.0 Volt**	**3.3 Volt**

Various notes about CPU's

Mixed notes.

Chip errors

The following miscalculations occur in 386, 486, and Pentium, when running Excel, Works, or Pascal, with the numbers 49 and 187:



All CPU's have faulty instructions. Recently faults have been discovered within the Pentium II and Cyrix 6x86MX.

The Pentium scandal

Pentium was hit by a scandal in late 1994, when an error in the mathematical co-processor (FPU) became publicly known. It simply miscalculated at a given division. Intel knew of the error from early that summer but more or less kept it secret.

Intel insisted that the error would occur extremely rarely. Compaq immediately modified their production to disable the FPU. Shortly thereafter, IBM announced that they would stop the production of Pentium based PC's. IBM had calculated that the error would occur every 24 days. At the time, IBM was working to extricate themselves from the Intel CPU monopoly. They were moving towards Power PC, Cyrix, and NexGen based PC's. Thus the scandal played right into their hands. You see the error here, where A3 should be equal to A1:

A3 =A1/B1*B1

	A	B	C
1	4195835	3145727	
2			
3	4195579		

Intel underestimated the significance of the miscalculations, certainly regarding users employing complex mathematical calculations. IBM over-dramatized the error for political reasons. This all happened in December 1994, while Intel was running their big TV campaign for Pentium.

That gave birth to a number of jokes: How many Pentium programmers are needed to screw in a bulb? (answer: 1,9990427). Why is Pentium not named 586? Because it would have to be called 585.999983405! In a different vein: How many Apple employees does it require to change a bulb? 7! One to hold the bulb and 6 to design T-shirts. And: how many IBM employees does it require to change a bulb? None! IBM simply announces a new feature called "black bulb."

Chip production

It takes a long time to manufacture a CPU. 3 to 5 million transistors must be placed on a tiny silicon wafer. Actually, it requires 90 24 hours round-the-clock workdays to produce a Pentium CPU.

CPU's are manufactured in large wafers containing 140 to 150 CPU's. Usually 110 to 120 of these perform perfectly. The rest are discarded. The wafers are burned, etched, and treated in hour long processes - layer by layer. In the CPU, there are up to 20 layers of silicon wafers with micro transistors.

CPU	**wafer size**
486	**.35 x .35 inch**
Pentium P54C	**.5 x .5 inch**
Cyrix 6X86	**.8 x .76 inch**

Moores Law

The CPUs have doubled their calculating capacity every 18 month. This is called "Moores Law" and was predicted in 1965 by Gordon Moore. He was right for more than 30 years. The latest CPUs uses internal wiring only 0.25 microns wide (1/400 of a human hair). But if Mores Law has to valid into the next century, more transistors has to be squeezed onto silicon layers.

And now there is a new hope. IBM has for the first time succeeded making copper conductors instead of aluminum. Copper is cheaper and faster, but the problem has to isolate it from the silicon. The problem has been solved with a new type of coating, and now chips can be designed with 0.20 micron technology. The technology is expected later to work with just 0.05 micron wiring!

Intel Owner's Club site

This site is a must if you are interested in the CPU's. The Intel Owner's Club is a free, easy way for members to:

- get the scoop on the latest Intel technologies
- get info on hot new software and technologies
- interact with Intel & technology experts
- download free software and games
- enter contests.

My membership has helped me to learn how to use the Intel web site, which holds a lot of information.

Now, let us see how CPU speed clocking works. Click for Module 3d.

To overview

Click & Learn. Module 3d. WWW.MKDATA.DK.

About over clocking

Introduction

All Pentium CPU's run with clock doubling. That's the way they are built. The PC works with two frequencies, which the user can adjust. The clock doubling is set on small *jumpers* on the motherboard. You simply set a clock doubling factor, to make the CPU work – but who says that you must use the factor listed in the manual?

If you are brave, you try to set your CPU to run *faster* than is is designed to run. Often it works. If you "cheat" the CPU in this manner to work faster, it is called *over clocking.* Over clocking is kind of a PC tuning, which can be fun to fool with – if you are interested in the technicalities of PC-hardware. Otherwise - skip it!.

If you are lucky, you can make a medium speed CPU run as fast as the top of the line version! Please note, I accept no responsibility for the result of your experiments. I will now try to explain the technologies in the over-clocking phenomenon. The interesting part is that, like much of the theory I tried to describe in in the modules 3a, 3 b and 3c, it all comes together here in the clock doubling technology.

By he way, much of the material is from Toms Hardware Guide, where I will refer you for further studies.

What is clocking?

The CPU works on two frequencies: An *internal* and an *external.*

- The external clock frequency (the bus frequency) is the speed between the CPU and RAM. In the Pentium CPU's it is actually the speed between L1 and L2 cache. In the Pentium II it is the speed between L2 cache and RAM.
- The internal clock frequency is the speed inside the CPU, that is between L1 cache and the various CPU registers.

For practical reasons you let these two frequencies depend on each other. In practice you choose a given bus frequency (between 60 and 100 MHz) and double it up a number of times (between 1½ and 5). The latter frequency become the CPU internal work frequency.

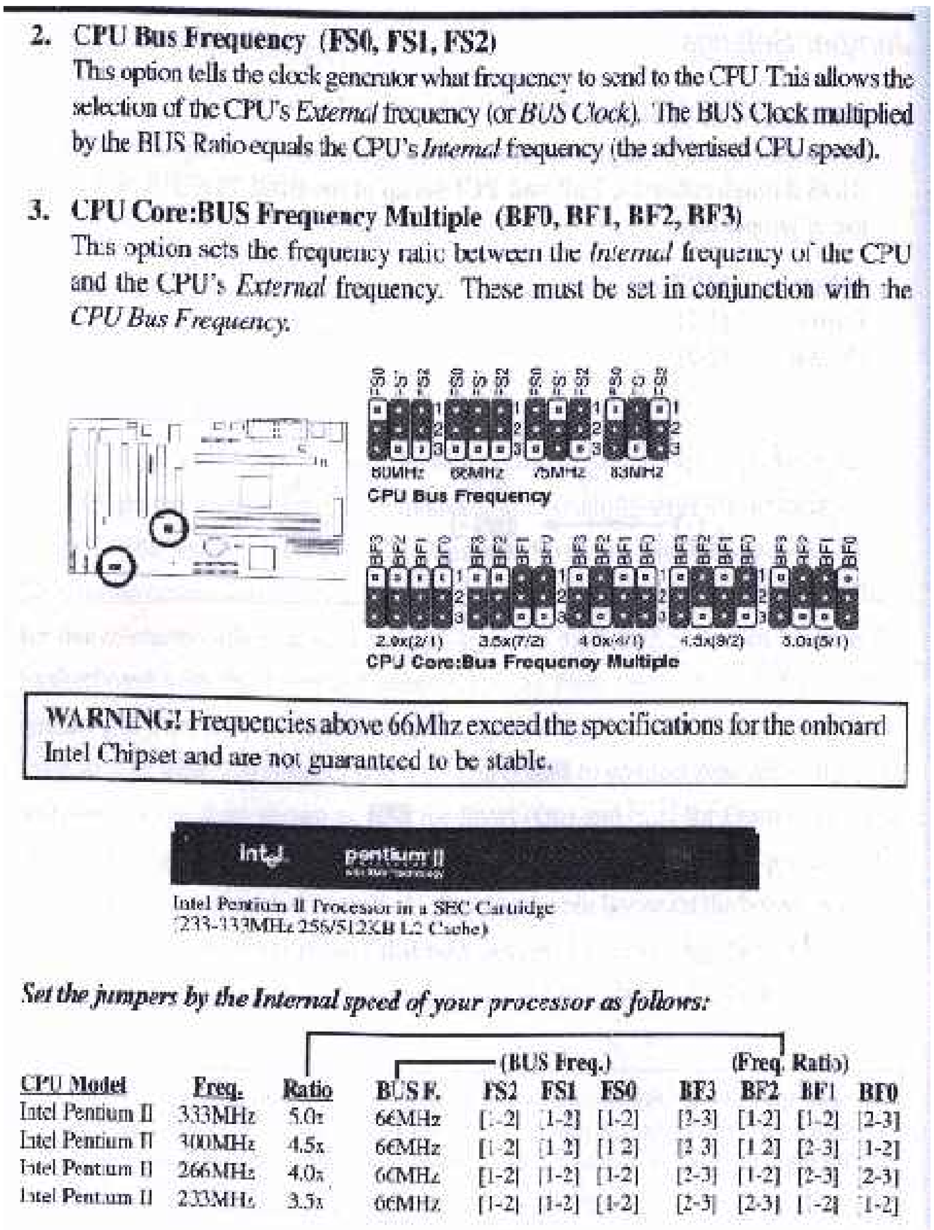
Here I show a number of *theoretical* CPU frequencies, resulting form different clock doublings: Many of these frequencies will actually never be used, but they are *possible* because of the system structure:

Bus frequencies	Clock doubling factors	Resulting CPU frequencies

60 MHz 66 MHz 75 MHz 83 MHz 100 MHz	1½, 2, 2½, 3, 3½, 4, 4½, 5	90 MHz, 100 MHz, 120 MHz 133 MHz, 150 MHz, 166 MHz, 180 MHz, 200 MHz, 210 MHz, 225 MHz, 233 MHz, 240 MHz, 250 MHz, 262 MHz, 266 MHz, 290 MHz, 300 MHz, 333 MHz 350 MHz, 375 MHz, 415 MHz, 450 MHz, 500 MHz

Note an important point: The CPU frequency is the result of the the bus frequency multiplied with a factor. If you increase the bus frequency, it affects the CPU frequency, which is also increased.

Look here at a page from the manual to a ASUS P2L97 motherboard. It has a clear instruction about how to set the the two values (bus frequency and clock factor). This motherboard accepts bus frequencies up to 83 MHz with a clock factor up to 5:



2. CPU Bus Frequency (FS0, FS1, FS2)
This option tells the clock generator what frequency to send to the CPU. This allows the selection of the CPU's *External* frequency (or *BUS Clock*). The BUS Clock multiplied by the BUS Ratio equals the CPU's *Internal* frequency (the advertised CPU speed).

3. CPU Core:BUS Frequency Multiple (BF0, BF1, BF2, BF3)
This option sets the frequency ratio between the *Internal* frequency of the CPU and the CPU's *External* frequency. These must be set in conjunction with the *CPU Bus Frequency.*

WARNING! Frequencies above 66Mhz exceed the specifications for the onboard Intel Chipset and are not guaranteed to be stable.

Intel Pentium II Processor in a SEC Cartridge
(233-333MHz 256/512KB L2 Cache)

Set the jumpers by the Internal speed of your processor as follows:

CPU Model	Freq.	Ratio	BUS F.	FS2 (BUS Freq.)	FS1	FS0	BF3 (Freq. Ratio)	BF2	BF1	BF0
Intel Pentium II	333MHz	5.0x	66MHz	[1-2]	[1-2]	[1-2]	[2-3]	[1-2]	[1-2]	[2-3]
Intel Pentium II	300MHz	4.5x	66MHz	[1-2]	[1-2]	[1-2]	[2-3]	[1-2]	[2-3]	[1-2]
Intel Pentium II	266MHz	4.0x	66MHz	[1-2]	[1-2]	[1-2]	[2-3]	[1-2]	[2-3]	[2-3]
Intel Pentium II	233MHz	3.5x	66MHz	[1-2]	[1-2]	[1-2]	[2-3]	[2-3]	[1-2]	[1-2]

What is over-clocking?

Since clock doubling and bus speed can be freely adjusted on the motherboard according to your desires, you can in principle make the CPU run at 300 MHz. You set the bus to 75 MHz and the clock factor to 4. Then the CPU runs at 300 MHz – if it runs. The question is whether the chip will tolerate that - if it will give a stable performance, since clock doubling means more than added heat.

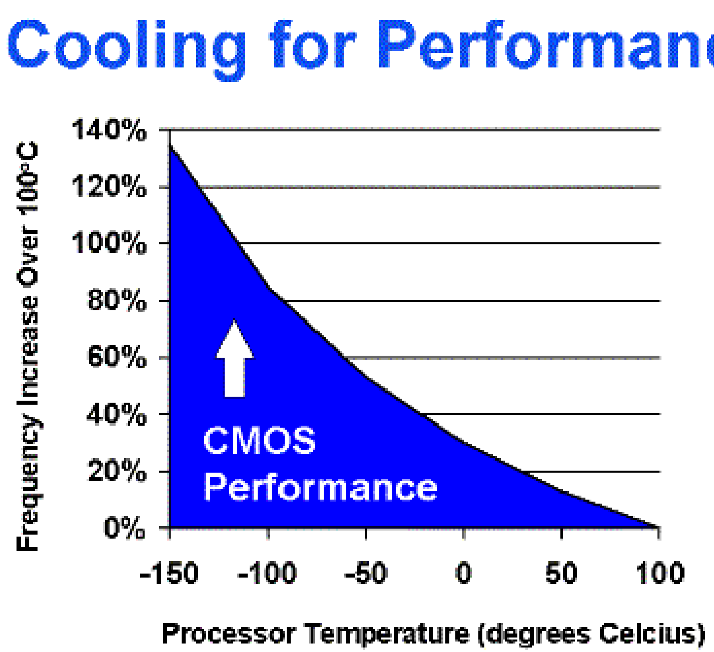
We have now seen that there are two frequencies which can be manipulated, if you want to re-clock the CPU:

- **The bus frequency can be increased**, let's say from 66 to 75 MHz.
- **The CPU frequency can be increased.** That can happen as a result of an increased bus speed, which also affects the CPU frequency, or it can happen by using a greater clock factor. Or it can happen through a combination of both techniques.

Both techniques result in a faster PC. If the bus frequency is increased, it affects all data transport to and from RAM. It will work faster, to the joy of all work done on the PC. When the CPU internal frequency is increased, many applications will be happily affected.

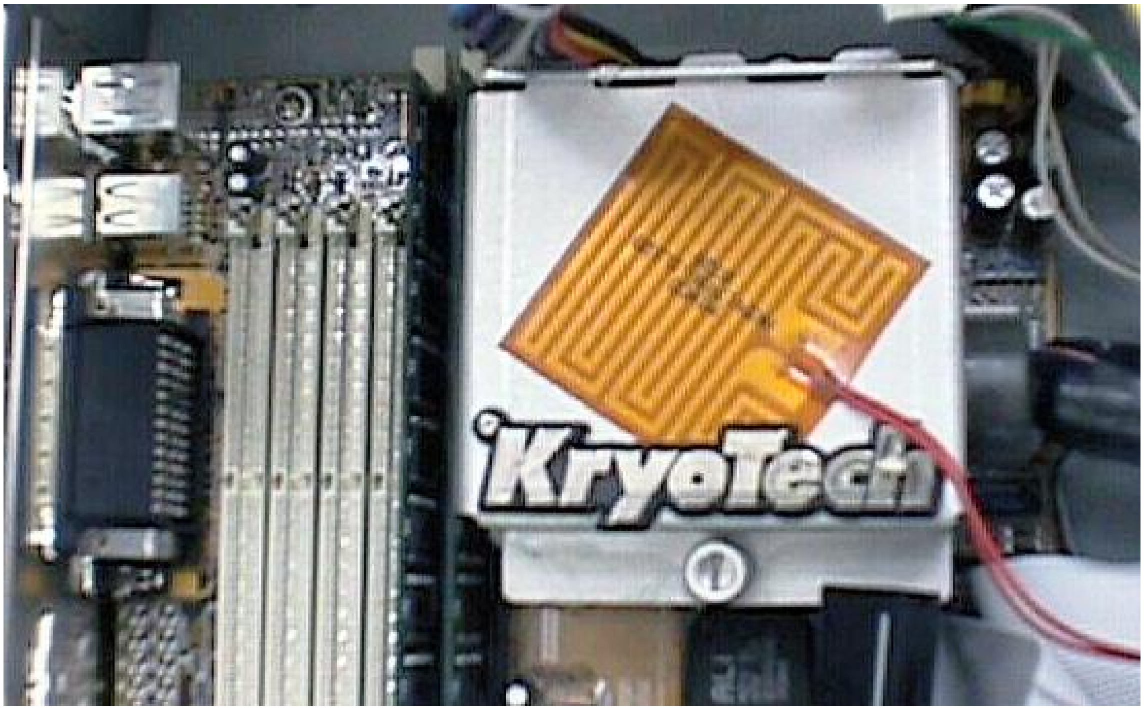
Cooling

The tuning will often work, but it requires good cooling of the CPU, the more cooling the higher you can have the clock frequency. CPU's are built in CMOS technology. That is a type chip which works better the cooler it is. See this relationship between temperature and performance:



You can see that the performance drops drastically with increased CPU temperature.

This problem caused the Kryotech company to manufacture coolers utilizing the Danish Danfoss compressors, just like in refrigerators. See this cooling unit on a CPU:



It is fed from the compressor in the bottom of the cabinet:

This form of cooling is extreme, but it works. Kryotech can make a standard CPU work at 400-700 MHz! But is requires that it is kept constantly cooled to -40 degree F.

If you like, look at Kryotech's Home Page http://www.kryotech.com/

This was to demonstrate that over-clocking can be a serious issue....

Which CPU's can be over-clocked?

The first CPU's which were dramatically over-clocked were AMD's 5x86 series. That was a 486 CPU, which could be forced up to an excellent performance at 160 MHz.

Since then especially Intel's Pentium CPU's have been over-clocked. Many of those seem to be sold with specs far from their optimum performance. Actually it is so easy resulting in many P133's were sold in 1996 as fake P166's. They worked fine, and the users did not know it. But Pentium MMX and Pentium II can also be re-clocked.

It appears at Intel is aware of this activity, and they don't seem to care. Unfortunately their CPU's come in two groups:

- Clock doubling works.
- Clock doubling does not work - it is *disabled* by the manufacturer.

Thus I can not guarantee that it always will work. But let me show a couple of examples, which I have made work with great results:

CPU	**Manufacturers spec**	**Tuning result**
Intel Pentium	2½ X 60 MHz = 150 MHz	3 X 66 MHz = 200 MHz
Intel Pentium Pro	3 X 66 MHz = 200 MHz	3½ X 66 MHz = 233 MHz
Intel Pentium II	3½ X 66 MHz = 233 MHz	4 X 75 MHz = 300 MHz

Looking at the three examples, number 1 and 3 show the best results, where both bus frequency and clock factors are increased. That simply moved the CPU up one class in performance.

Here is a table of the clock factors, which the CPU's theoretically can accept (according to my studies):

CPU	**Clock factor**
Intel Pentium (P54C)	**1½, 2, 2½, 3**
Intel Pentium Pro	**2½, 3, 3½, 4**
Cyrix 6x86	**2, 3**
Cyrix 6x86MX (M2)	**2, 2½, 3, 3½**
Intel Pentium MMX (P55C)	**2, 2½, 3, 3½**
AMD K5 PR75 - PR133	**1½**
AMD K5 PR150 and PR166	**2**
AMD K6	**2, 2½ 3, 3½**
Intel Pentium II	**2, 2½, 3, 3½, 4, 4½, 5**

AMD and Cyrix chips are special, in that they do not always respond to motherboard settings. It is like they determine their own frequencies.

Which disadvantages and risks are there in over-clocking?

Many factors need to be considered, when you start tampering with these system settings. Watch out for:

- Heat. Can the CPU dissipate the heat?
- The L2 cache RAM of Pentium II cartridges - how fast can it work?
- RAM speed. Can it keep up with the system bus?

- The I/O bus. Can PCI and EIDE units keep up?
- Will the software still work?

The last two problems are associated with increased system bus speed. This kind of over-clocking gives the best results. However those also create the biggest problems, at least in my experience.

The CPU gets hot

First of all the higher CPU frequency causes more wear on the chip. It is said that a CPU can last 10 years. However do not count on that if you over-clock it. Actually I am less concerned about the wear. Of course you should not allow the the chip to over heat, but I have never heard about burnt out CPU's. In news groups you can read about various monster fans used to cooling of totally over-clocked CPU's.

RAM speed

Another problem is in the relationship with the bus frequency. Here we are talking about the system bus, which connects RAM with the CPU. If you increase this speed, RAM must be able to keep up. Here is a guideline table for the maximum bus frequencies with different RAM types:

RAM type	Speed	Maximum bus frequency
FPM	60 ns	66 MHz
EDO	50 ns	75 MHz
SD	10 ns	100 MHz

I/O speed and experiments with that

The area which has given me the most problems is the increased PCI speed. The PCI bus runs at half the system bus frequency. Thus if we increase it, it also affects the PCI bus:

System bus speed	Resulting PCI speed
66 MHz	33 MHz
75 MHz	37.5 MHz

When we increase the PCI bus speed, a number of units are affected. They may not always agree with the faster pace. This includes:

- The EIDE hard disk
- The video card
- The network controller and other I/O cards.

I have experienced some hard disks, which could not tolerate a bus frequency of 75 MHz, and there is nothing you can do about that. Lately I have experimented with my Pentium II, which was bought as a 233 MHz model.

First I made it run at 3½ X 75 MHz. It worked fine with CPU, RAM (10 ns SD) and hard disk (IBM DHEA). But the net card (a cheap 10/100 Ethernet card) refused. When I copied large volumes of files on the net, it froze up - stopped. It was quite obvious that the problem was in the net card.

I had to accept the traditional 66 MHz. But to soothe the pain, it turned out to run excellently with a clock factor of 4 - thus at 266 MHz.

Within a couple of weeks I was in the mood to experiment again. I now found an adjustment in the setup program. It is called PCI latency. It is not explained anywhere, but it has a default value of 32. I increased it to 36 and increased the bus frequency to 75 MHz – it works. Now the net card runs without problems.

Then I hoped to speed the system bus up to 83 MHz, which should give a significant performance improvement for all RAM transport. My 10 ns SD RAM can certainly handle 83 MHz. But no, it did not work. Regardless of the PCI latency, the PC would not start. This indicates that the PCI latency setting does not work like I expected. Maybe it has nothing to with this - I do not know.

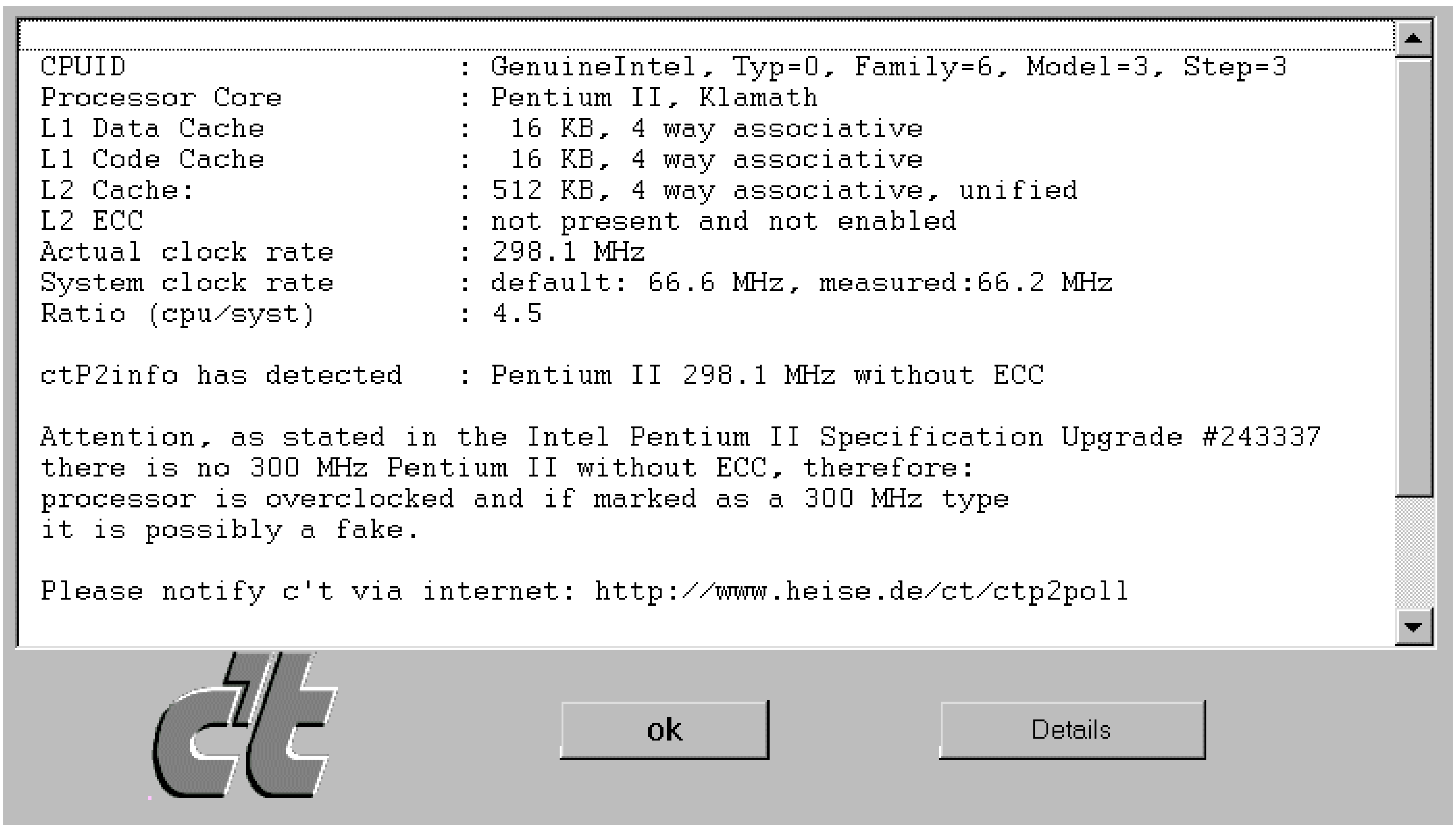
My explanation is, that the video card could not tolerate the 41.5 MHz PCI frequency. Nothing appeared on the screen.

Now the PC runs fine at 4 X 75 = 300 MHz. There can be an occasional unexplained break-down in Windows 95 (that happens under other circumstances also), which I blame on the drastic over-clocking. However, the advantages of the significant performance improvement far exceed the annoyance of these small interruptions, which happen far from daily.

Windows NT 4.0 does not install with over clocked CPU. The program tests for "genuine Intel", and seems to register the change in clock frequency. And then it will not work. But if you install NT first, then you can over clock afterwards and NT will work. Actually NT is quite sensitive. One of my friends experienced some peculiar errors. The solution turned out to be moving the RAM module from one socket to another!

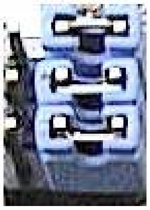
Fake Pentium II's

Since some Pentium II-233 perform very well at 300 MHZ, they have been sold as such ones. To test your own Pentium II, you can download this test program from C't, which can check your Pentium-II. Here is the interface of the Windows 95 version, which correctly detected my CPU to be overclocked:



Jumpers on the system board

To set the clock-doubling, some small switches (called jumpers) have to be reset. They are located on the motherboard, as you see here:



You can read in the system board manual how to set them. Or you can look at the system board! In the picture below you can see some of the printed information *on* the motherboard (this is an ASUS TX97 with socket7).

Here you can read which jumpers to set to select clock doubling 1, 1½, 2, 2½, 3, 3½ and 4 for 6 types of processors:

- P54C and K5
- P55C, K6 and M2 (Cyrix 6x86MX)
- M1 (Cyrix 6x86)

FREQ RATIO BF0	BF1	P54C K5	P55C K6/M2	M1
1-2	1-2	X1.5	X3.5	X3
2-3	1-2	X2	X2	X2
1-2	2-3	X3	X3	X4
2-3	2-3	X2.5	X2.5	X1

Last revised: 8 Jun 1998. [To Overview.]

Click & Learn. Module 4a. WWW.MKDATA.DK

Drives are storage media's

A drive is the name name for several types of *storage media.* There are also storage media, which are not drives (RAM, Tape Streamers), but on these pages, we will discuss the drives.

Common to drive media is:

- A *file system* can be assigned to them.
- They are recognized by the operating system and they are assigned a *drive letter*.

During start up, drives are typically recognized by the PC system software (ROM-BIOS + operating system). Thus, the PC knows which drives are installed. At the end of this configuration, the appropriate drive letter is identified with each drive. If a drive is not "seen" during start up, if will not be accessible to the operating system. However, some external drives contain special soft-ware, allowing them to be connected during operation.

Here some examples of drives:

Storage media	Drive letter
Floppy disks	A: B:
Hard disk	C: D: E:
CD ROM	F:
MO drive	G:
Network drive	M:
RAM disk	O:

On this and the following pages, I will describe the various drive types, their history and technology. The last two drive types in the above table will not be covered.

Storage principles

Storage: Magnetic or optic. Data on any drive are *digitized.* That means that they are expressed as myriad's of 0's and 1's. However, the storage of these bits is done in any of three principles:

The physical drive principle	Disk types

Magnetic	**Floppy disks** **Hard disk** **Syquest disks**
Optic	**CD ROM** **DVD**
Magneto optic	**Zip drive** **LS-120 disks** **and others**

Interface

Individual drives are connected to other PC components through an interface. The hard disk interface is either IDE or SCSI, which in modern PC's is connected to the PCI bus. Certain drives can also be connected through a parallel port or the floppy controller:

Interface	**Drive**
IDE and EIDE	**Hard disks (currently up to 8 GB)** **CD ROM**
SCSI	**Hard disks (all sizes) and CD ROM**
ISA (internet)	**Floppy drives** **CD ROM an MO drives connected through parallel port**

Let us start evaluating the drives from the easy side:

Floppy drives

We all know diskettes. Small flat disks, irritatingly slow and with too limited storage capacity. Yet, we cannot live without them. Very few PC's are without a floppy drive.

Diskettes were developed as a low cost alternative to hard disks. In the 60's and 70's, when hard disk prices were exorbitant, It was unthinkable to use them in anything but mainframe and mini computers.

The first diskettes were introduced in 1971. They were 8" diameter plastic disks with a magnetic coating, enclosed in a cardboard case. The had a capacity of one megabyte. The diskettes are placed in a *drive*, which has read and write heads. Conversely to hard disks, the heads actually *touch* the disk, like in a cassette or video player. This wears the media.

Later, in 1976, 5.25" diskettes were introduced. They had far less capacity (only 160 KB to begin with). However, they were inexpensive and easy to work with. For many years, they were the standard in PC's. Like the 8" diskettes, the 5.25" were soft and flexible. Therefore, they were named *floppy disks*.

In 1987 IBM's revolutionary PS/2 PC's were introduced and with them the 3½" hard diskettes we know today. These diskettes have a thinner magnetic coating, allowing more tracks on a smaller surface. The track density is measured in TPI (tracks per inch). The TPI has been increased from 48 to 96 and now 135 in the 3.5" diskettes.

Here you see the standard PC diskette configurations:

Diskette size	**Name**	**Tracks per side**	**Number of sectors per tracks**	**Capacity**
5.25" Single side	**SD8**	**40**	**8**	**40 X 8 X 512 bytes = 160 KB**
5.25" Double side	**DD9**	**40**	**9**	**2 X 40 X 9 X 512 bytes = 360 KB**
5.25" Double side High Density	**DQ15**	**80**	**15**	**2 X 80 X 15 X 512 bytes = 1,2 MB**
3.5" DD	**DQ9**	**80**	**9**	**2 X 80 X 9 X 512 bytes = 720 KB**
3.5" HD	**DQ18**	**80**	**18**	**2 X 80 X 18 X 512 bytes = 1,44 MB**
3.5" XD (IBM only)	**DG36**	**80**	**36**	**2 X 80 X 36 X 512 bytes = 2,88 MB**

Diskette drives turn at 300 RPM. That results in an average search time (½ revolution) of 100 ms.

The floppy controller

All diskette drives are governed by a controller. The original PC controller was named NEC

PD765. Today, it is included in the chip set, but functions like a 765. It is a programmable chip. It can be programmed to handle all the various floppy drive types: 5.25" or 3.5" drives, DD or HD etc.

The controller has to be programmed at each start up. It must be told which drives to control. This programming is performed by the start up programs in ROM (read module 2a). So you don't have to identify available drive types at each start up, these drive parameters are saved in CMOS RAM.

The floppy controller reads data from the diskette media in serial mode (one bit at a time. like from hard disks). Data are delivered in parallel mode (16 bits at a time) to RAM via a DMA channel. Thus, the the drives should be able to operate without CPU supervision. However, in reality this does not always work. Data transfer from a diskette drive can delay and sometimes freeze the whole PC, so no other operations can be performed simultaneously.

To continue:

Read Module 4b about hard disks.

Read Module 4c about optical media's (CD-ROM and DVD).

Read Module 4d about MO drives.

Read Module 4e about tape streamers (which *are not* drives).

Read Module 5c about SCSI.

Read Module 6a about file systems.

To overview

Click & Learn. Module 4b. WWW.MKDATA.DK

Hard disks

Hard disks consist of one or more magnetic disks contained in a box. They are used as storage media in the PC, where you store programs and other digital data.

The magnetic storage hard disk is based on a 40 year old technology. It has been and still is being improved rapidly. Hard disks continue to shrink in size, gain increased storage capacity and increased transfer speeds. The development has been tremendous during the last 10 years. Indications are that this will continue for a long time.

When buying a PC, It is a good rule to include a large and fast hard disk. You can never buy too large a hard disk and the data transfer speed is decisive for the PC's performance.

An evaluation of the hard disk, its configuration and performance, involves several different technologies. That is the subject of this page:

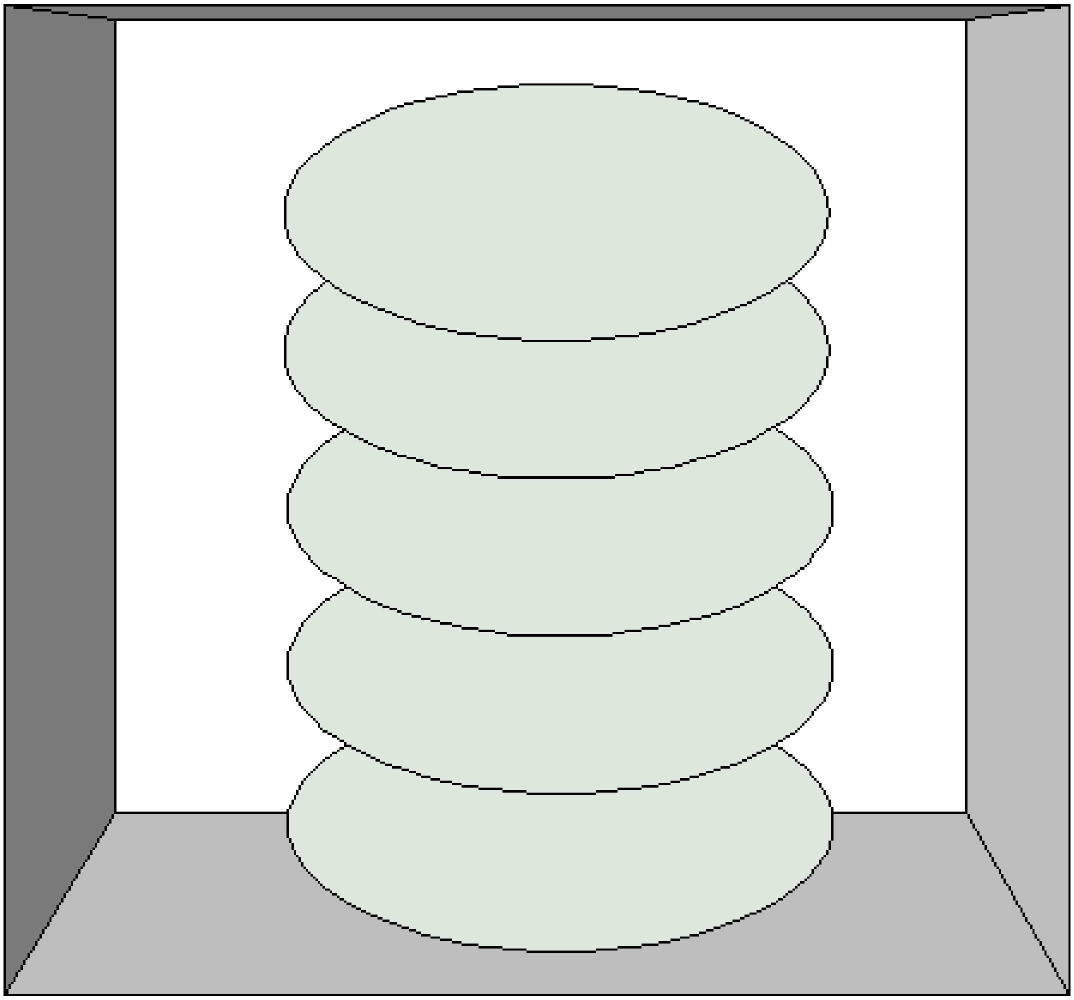
- ***The mechanical disk.*** The physical disk construction, RPM, read/write head, data density, etc.
- ***The cache.*** The hard disk always has some cache RAM onboard. It serves as a buffer, so the data being physically read is best utilized.
- ***The interface.*** The connection between the hard disk and other PC components. That is called *interface* - the connection to a data bus, the controller principle.
- ***Formatting etc.*** Disk formatting, control system, cache, etc. I presume you are running in Windows 95, which has the best access to the hard disk.

I want to illustrate the inter-action between these features, thus giving a comprehensive picture of the hard disk and its technologies.

Historic

First, let us look at the hard disk history. IBM introduced the first hard disk in 1957. That was a major project. It consisted of 50 platters, 24 inch diameter, with a capacity of 5 MB, a huge storage media for its time. It cost $35,000 annually in leasing fees (IBM would not sell it outright). The first model to use "float on air" technology for the read/write heads was named Winchester 3030. So named because it was developed in Winchester, England and had two sides, each of which could store 30 MB. To some people, this designation was reminiscent of the famous Winchester 3030 repeating rifle.

Later, the disk platters shrunk to 14" and 8" diameter. They were installed in towers containing dozens of these magnetic platters.



In the early years of PC development, the low cost floppy drives were the preferred storage media. But with IBM's XT in 1983-84, the hard disk became the preferred media. The first hard disks were rather large units (5.25" diameter) and of poor quality. I have replaced numerous 5, 10 and 20 MB hard disks during 1986-88, since these early PC hard disks had an incredible short life span. Since then they have improved a lot.

The modern hard disks are 3.5" diameter. A typical example is the Quantum Fireball, which you see above. The cover plate has been removed, so you can see the top arm with its read/write head.

Hard disks can be found in much smaller sizes (all the way down to match box size). However, for ordinary, stationary PC's the 3.5" is the best. They are inexpensive to manufacture, and they are faster.

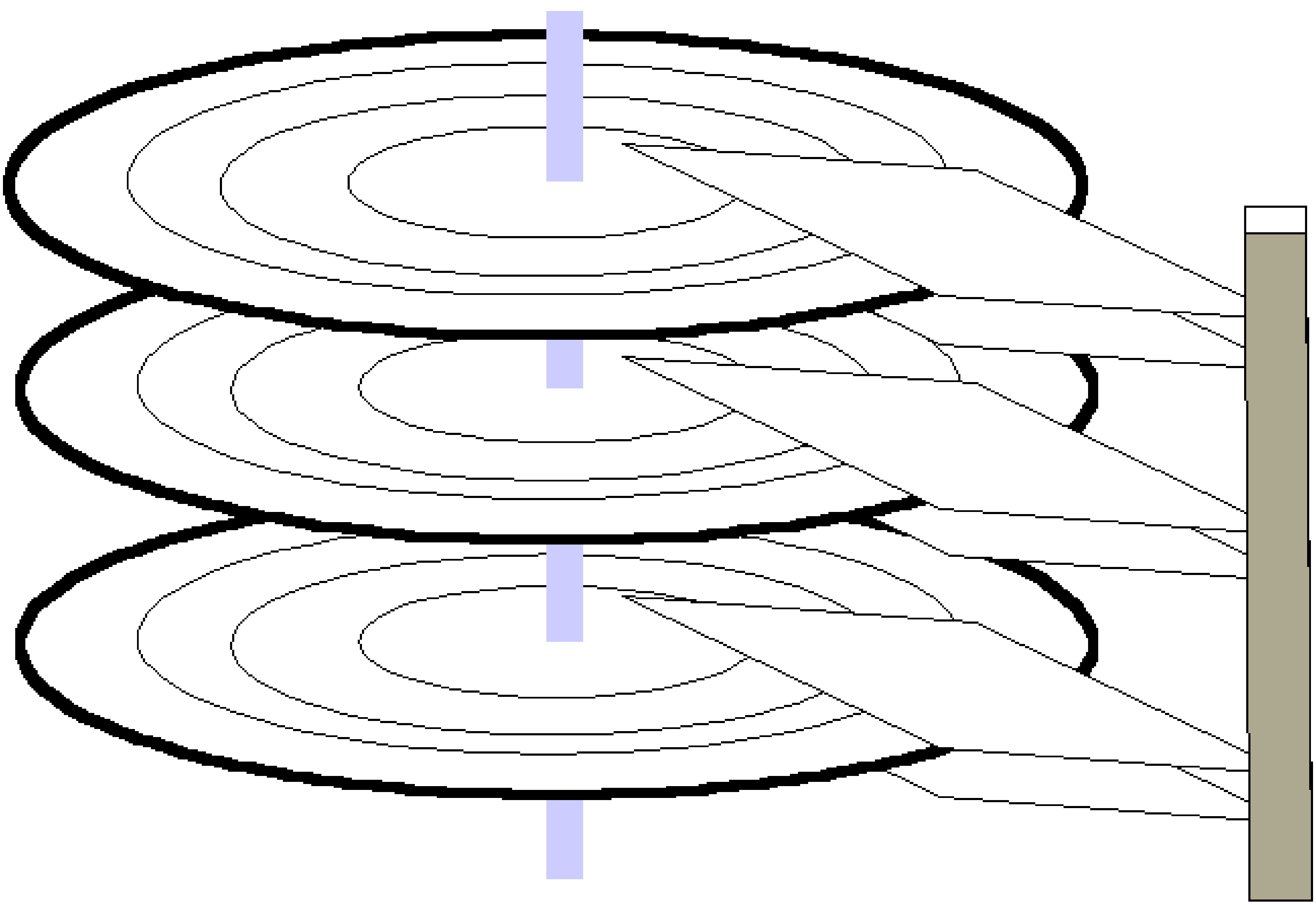
Physical aspects of the hard disk

First, let us look at the construction of the hard disk.

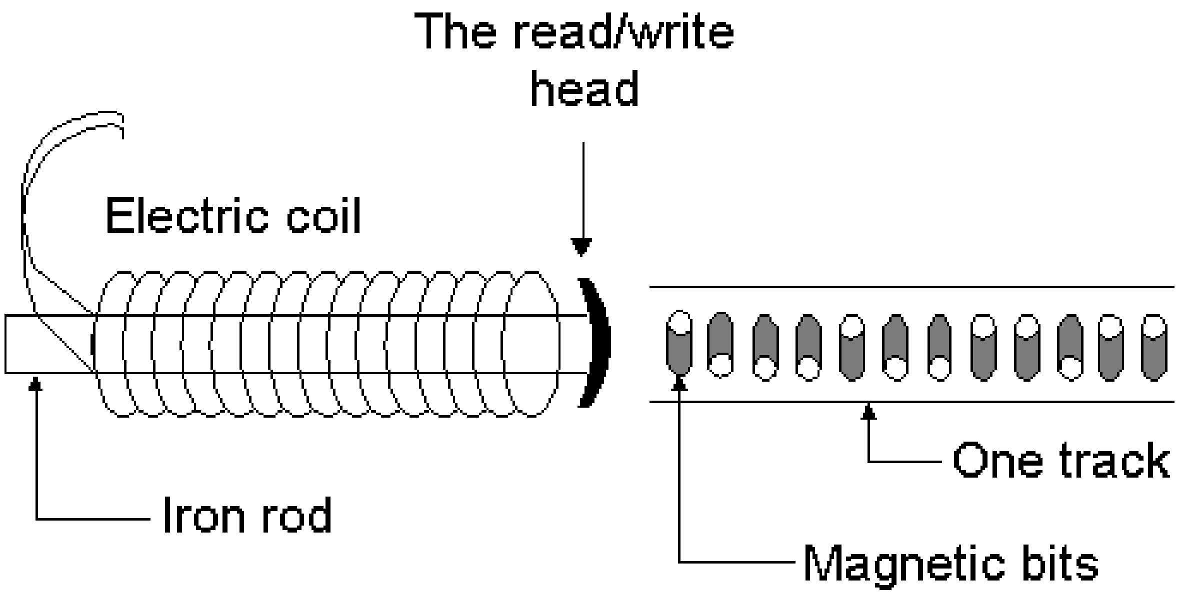
Read/write heads

All hard disks consist of thin platters with a magnetic coating. They rotate quite fast inside a metal container. Their design causes them to ride on a microscopic cushion of air, without touching the platter. Data are written and read by read/write heads. They register bits from the magnetic coating, which races past them. On the illustration below, you see a hard disk with three platters. It has 6 read/write heads, which move synchronously.

The arms, which guide the movement of the read/write heads, move in and out almost like the pick-up arm in an old fashioned phonograph. As illustrated below, there will typically be 6 arms, each with read/write heads. The synchronous movement of the these arms is performed by an electro-mechanical system called head actuator. The hard disk data can only be attained via one head at a time.



The read/write head consists of a tiny electromagnet. The shape of the head end acts like an air foil, lifting the read/write read slightly above the spinning disk below.



When the disk rotates under the read/write head, it can either read existing data or write new ones:

- If a current is applied to the coil, the head will become magnetic. This magnetism will orient the micro magnets in the track. This is write mode.
- If the head moves along the track without current applied to the coil, it will sense the micro magnets in the track. This magnetism will induce a current in the coil. These flashes of current represent the data on the disk. This is read mode.

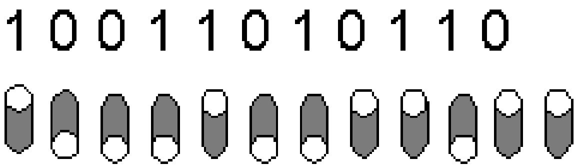
The read/write heads are by far the most expensive parts of the hard disk. They are incredibly tiny. In modern hard disks they float between 5 and 12 micro inches (millionths of an inch) above the disk. When the PC is shut down, they are auto parked on a designated area of the disk, so they will not be damaged during transport.

Domains

The bits are stored in microscopic magnets (called *domains*) on the disk. They are written in this manner: Before recording data, the drive uses the read/write heads to orient the domains in a small region so that the magnetic poles all point in the same direction. Then:

- A reversal of polarity is interpreted as a digit one.
- Unchanged polarity is interpreted as a digit zero.

If we read the magnets from right to left, we might see the following example:



The magnetic disks

The magnetic disks are typically made of aluminum. There are also experiments with disks made of glass. The disks are covered with an ultra thin magnetic coating. With improved coating technologies, an increasing number of micro magnets can be placed on the disk. Currently, there more than 2000 tracks per inch disk radius. There are only 135 on a floppy disk. The narrower the tracks are, the bigger the disk capacity gets. At the same time the magnetic signals get weaker and weaker. Therefore, the read/write heads must get closer to the disk. This requires even smoother platters, etc.

An other improvement in modern disks is the employment of a technology called Multiple Zone Recording. This allows for about twice as many sectors (120) in the outermost track as in the innermost. Thus, outer tracks, which are much longer, can hold much more data than inner tracks. Previously, all tracks had the same number of sectors, which was not very efficient.

Writing in layers

Since a hard disk typically contains three platters with a total of 6 read/write heads, the concept *cylinders* is employed. Read/write heads move synchronously. Therefore, data are written up and down from platter to platter. Thus, one file can easily be spread over all 6 platter sides. Let us say the writing starts on track 112 on the first platter. That is completed and writing continues on track 112 - only from read/write head number 2. Then it continues to numbers 3, 4, 5 and 6. Only then does writing move to track 113.

In this case, a cylinder consists of 6 tracks. For example, cylinder number 114 is made up of track number 114 on all 6 platter sides.

Development

Everyone wants faster, cheaper disks with increased capacity. Therefore, hard disk technology undergoes an explosive development. There are two major trends in this development:

- Data are packed increasingly closer with new coating and read/write techniques.
- The disks rotate faster and faster.

It is my impression that the various hard disk manufacturers alternately develop new, sophisticated technologies, which spread with the speed of lightning to other brands. A couple of years ago, IBM suddenly introduced lightning fast disks with new MR heads. Everyone uses those today. In 1997 Quantum introduced the Ultra DMA interface, which all other manufacturers now use as well.

Speed generates heat

To day, an ordinary "lazy" hard disk rotates at 4500 RPM. The better run at 5400 and the best at 7200, but there is much development in this area. One of problems is the heat generation, which increases with the speed. The fewer and smaller platters to turn, the faster they can reach their full speed. Another problem is noise. I had to return a disk, which turned at 7200 RPM with the sound of a dentist's drill. It was intolerable to be in the same room!

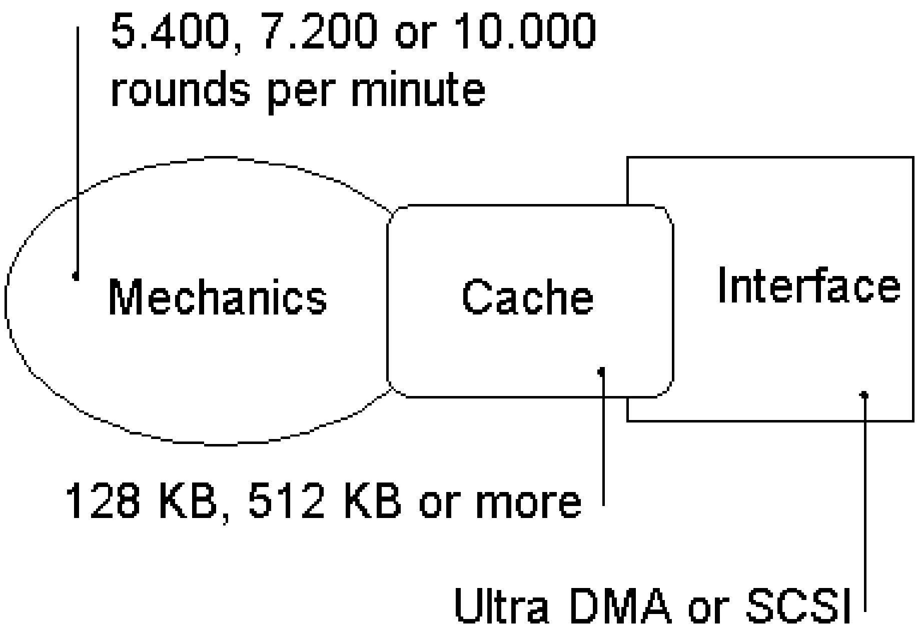
The best SCSI disks available rotates at 10.000 and 14.000 RPM. They are expensive but ultimately they provide the highest transfer capacity.

Controller principles

Further down this page is a review of the interfaces which are used for hard disks. But, first look at this brief summary:

Controller	**Description**
MFM, RLL and ESDI	Older standards for hard disk interface
IDE	Simple, primitive interface. Data are delivered on the ISA bus, resulting in low transfer speed. Disks < 528 MB. Is not used on new PC's.
EIDE	Improvement of IDE. Data are delivered on the PCI bus with Bus Master control, which results in high transfer speed. Large disks (currently up to 8,5 GB). Room for four units, which are typically connected directly to the system board. Inexpensive and effective disks. Ultra DMA is the best interface among the EIDE varieties.
SCSI	High end controller system, where the units are connected to a special (expensive) controller. The hard disks are generally of highest quality, fast and with a long life span. An ordinary SCSI controller can run 7 hard disks of 18 GB.

All together we have three levels of technology which all can be optimized for better hard disks:



Hard disk limitations

Up through the years, hard disks have suffered from a long list of software related limitations. Read about that under the file system.

Hard disk speeds

There are countless test programs and measuring methods to evaluate the various hard disks.

Don't place too much stock in the sales person's presentation of seek times. Many hard disks are advertised with a number like 8 ms. That refers to a seek time, which is measured in milliseconds. There are many different seek times. That makes comparison difficult. You can measure in terms of:

- Average track to track speed. How long does it take the actuator to move read/write heads from one track to another? There are typically 3000 tracks on a platter side. There, a track change could be to just one over. That might take 2 ms. Or, it could be up to 2999 tracks over. That might take 20 ms. On current hard disks, the average seek time will be between 8 and 14 ms.
- Change time between read and write - That takes time also.
- Wait time for the right sector. When the arm moves to a track, it must wait for the right sector to appear under the head. That takes time also. On the average, the platter must rotate ½ revolution, to reach the right sector. This time is directly proportional to the disk rotation speed. On modern hard disks it usually is between four and eight ms.

When I test a hard disk, I emphasize practical applications. You can take a stopwatch and measure, for example how long it takes to start Windows 95 or Word 97 (possibly including a large file). That type of measurement can really tell you something about the hard disk's performance. However they must be under under comparable circumstances. System board, CPU, and the driver program also influence the results.

Hard disk interface

Read here in more detail about the different interface types.

MFM and RLL

MFM and RLL are actually coding principles for the hard disks. They are not interface standards. The coding occurs from controller to the hard disk. As a coding principle, RLL was more effective than was MFM, so in the "old days" you could experiment using RLL controllers to MFM disks.

WD 1003

MFM as well as RLL are WD1003 compatible, meaning that the standards would work with the at that time most widely used controller chip from Western Digital.

ST 506

ST 506 is an interface, which was used both with RLL and MFM. There is serial connection from controller to disk. The ST 506 controller functions as as a converter from the serial read/write head data to the 8 or 16 bit parallel bus. ST 506 was the most widespread controller standard before IDE.

IDE

Integrated Device Electronics. Under the IDE standard, the controller chip WD 1003 is mounted directly on the hard disk, not on the IDE adapter. This means that the conversion to parallel data is already done on the disk. Because of the short serial cable, this increases the transfer speed significantly relative to MFM and RLL. IDE is a simple adapter. The adapter itself contains only amplifying circuits to/from the I/O bus. Therefore it is inexpensive. The IDE controller does not care whether the hard disk works internally with MFM or RLL coding.

ESDI

ESDI is an improvement over the ST506 standard. An ESDI disk operates on a common 16-bit AT bus (ISA-bus), but it is better put together than an ordinary IDE. This results in an almost doubling of the transfer speed between hard disk and controller/bus. ESDI is also different in many other ways. Among the features are a sector on the hard disk which identifies its number of tracks, cylinders, etc. This information is usually stored in CMOS.

EIDE

EIDE is the current standard for low cost, high performance hard disks. EIDE stands for Enhanced IDE. That is precisely what it is. Some manufacturers call it ATA.

All Pentium system boards since 1995 have a built in EIDE controller. That allows the hard disk to be connected directly to the system board. The EIDE standard is substantially improved, relative to the old IDE. Here are some of the improvements:

- **The hard disk size** can exceed the 528 MB, which was the IDE limit.
- **The hard disk interface** is moved from the ISA bus to the high speed bus PCI.
- **Four units** can be connected on the mainboard, which has two EIDE channels. Each channel can be connected to a primary and a secondary unit.

Transfer speeds

With connection directly to the PCI bus, EIDE has transfer speeds and disk capacities which far exceed the older controller principles. EIDE exists in different editions, such as PIO 3, PIO 4 and Ultra DMA. The latter is the one to choose. The different PIO *modes* are significant for transfer speed. PIO 3 can transfer up to 13 MB/sec, while PIO 4 promises 16.6 MB/sec. UDMA promises up to smashing 33 MB/sec. These numbers are *theoretical* and they do not hold true in actual use. The fastest actual transfer speed you can experience from an EIDE disk will be 5-10 MB/sec. That is still good.

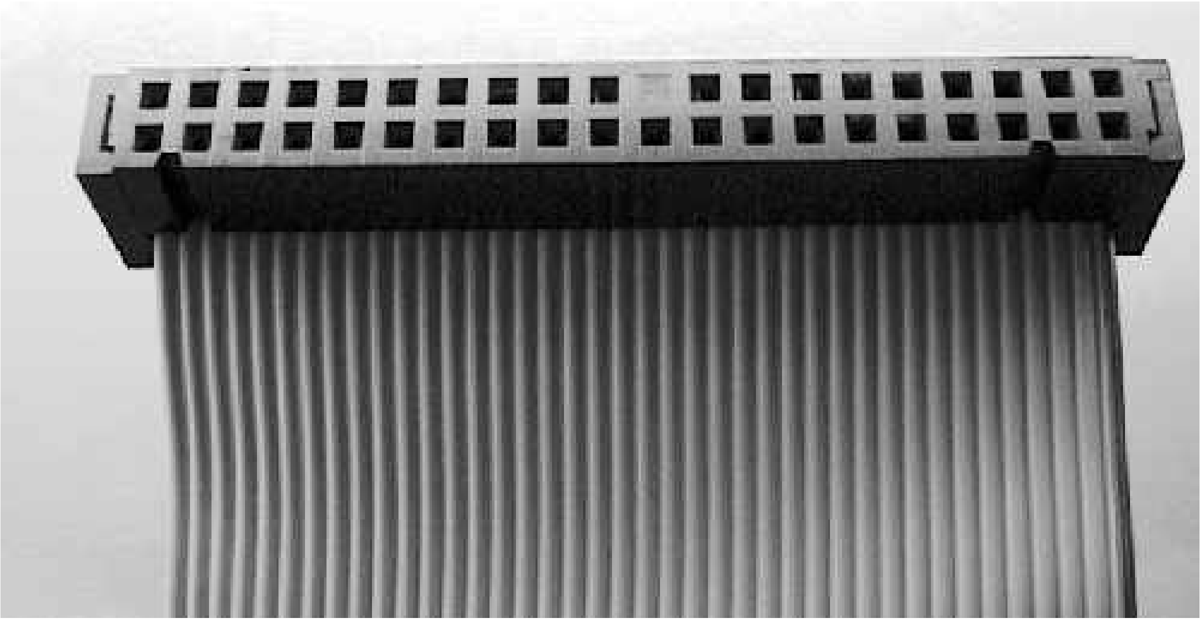
Four units

The EIDE interface is not only intended for the hard disks. There are four channels, which can be connected to four independent units:

- Hard disks (must be on the primary channels, which on some system boards have the greatest transfer capacity)
- CD ROM drives
- DVD drives
- LS 120 drives and other MO drives
- Tape streamers

EIDE is thus designed as an inexpensive all around interface, which can be connected to all kinds of storage media. It has a clever auto detect function, which often makes it possible to connect EIDE units such as hard disks directly to the system board and function immediately. The BIOS in the PC will find the necessary instructions about the drive via the auto detect function, and you need not make any adjustments in the CMOS Setup program, as was necessary with earlier IDE units.

Hers you see an EIDE cable:



There are four connections like this on the mainboard. Note the blocked hole in top center. Note also the stripe (which is red) on the right edge of the cable. It tells that lead number one is in that side. Both features prevent reverse cable installation.

Ultra DMA (under construction, more to come)

The latest improvement of EIDE is called Ultra DMA or Ultra ATA. It is a new interface, patented by Quantum. It is supported by all system board manufacturers.

The technique consists of an improvement in the interface - the guiding electronics, which deliver the hard disk data to the system board. Quantum succeeded in removing bottle necks in the existing guiding electronics to the EIDE disks. The hard disk speed in itself has not improved, but its data busses have been optimized.

Practical measurements show that Ultra DMA disks (Quantum ST or the very powerful IBM Deskstar 5) perform 15-25% better than similar PIO 4 disks. This is a clear improvement, even though the disks in no way can deliver the 33 MB/sec., as advertised.

To be able to get the the benefits of such a disk, the system board and with that the chip set must be prepared for Ultra DMA, so you can utilize such a disk. As always when you buy a new PC, check the chip set. It is important that it supports Ultra DMA, which provides a solid performance increase at no extra cost.

Click for information about he most advanced and elegant controller principle of all: *SCSI.*

To overview

Click & Learn. Module 4c. WWW.MKDATA.DK

Content:

The Optic Medias (CD ROM's and DVD)

CD ROM and DVD are *optic readable* media, contrary to hard disks, floppy disks and tapes, which are *magnetic*.

The optic storage media are read with a very thin and very precisely aimed laser beam. They supplement the magnetic media. They have clear advantages in the areas of data density and stability: Data can be packed much more densely in an optic media than in a magnetic media. And they have much longer life span. It is presumed that a magnetic media, such as a hard disk or DAT (*digital audio tape*) can maintain their data for a maximum of five years. The magnetism simply fades away in time. Conversely, the life span of optic media is counted in tens of years.

Let us take a closer look at these disks, which are becoming increasingly popular for all types of information, education and entertainment. There are different types:

The Compact Disk

The compact disk (CD) was introduced by Philips and Sony in 1980 to replace LP records. It is a small plastic disk with a reflecting metal coating, usually aluminum. Myriad's of tiny indentations are burned into this coating. These indentations contain the music in millions of *bits*. The CD is organized in tracks. Each track is assigned a number.

The big advantage of the CD is its high-quality music reproduction and total absence of back ground noise as well as a great dynamic. During operation, the software in the drive can correct errors caused by such things as finger marks on the disk. All in all, the CD is an excellent music storage media.

The CD-ROM

The CD-ROM (Read Only Memory) came as an extension of the CD in 1984. In principle, the media and the drives are the same. The difference is in the data storage organization. In a CD-ROM, the data are stored in sectors, which can be read independently - like from a hard disk.

The CD-ROM has become an important *media* in the PC world. It can hold 650 MB of data, and it is very inexpensive to produce. Today, there are three types of CD drives and DVD drives are on their way:

Drive type	Name	The drive can
CD-ROM	Compact Disk Read Only Memory	Read CD-ROM and CD-R

CD-ROM multiread	--"--	Read CD-ROM, CD-R and CD-E
CD-R	Compact Disk Recordable	Read CD-ROM and CD-R. *Write once* on special disks named CD R
CD-E or CD-RW	Compact Disk Erasable or Compact Disk ReWritable	Read CD-ROM's and CD-R Write and re-write on special disks (CD-E).
DVD RAM	Digital Versatile Disk Random Access Memory	Reads all CD formats. Reads DVD ROM. Reads and writes DVD disks

Let us start by look at the CD-ROM construction. To facilitate understanding, it will be easiest to compare it with other disk types, especially the hard disk. The CD-ROM is a plastic disk of 4.6" diameter.

It is placed in a CD-ROM drive, which is like a drawer in the PC cabinet:

When the CD-ROM disk is placed in the drive, it starts to spin the disk. It reaches operating speed in one to two seconds. Then the drive is ready to read from the disk.

Drives and operating system

The drive must be assigned *a drive* letter. That is a task for the operating system, which must be able to recognize the CD-ROM drive. That is usually no problem in Windows 95. However, the alphabet can be quite messy, if there are many different drives attached. Each drive must have its own letter. They are assigned on a first come first-serve-basis.

The CD-ROM drive usually gets the first vacant letter after other existing drives, typically D, E, or F. But the letter can be changed.

Once the CD-ROM spins and the operating system (DOS or Windows) has "found" the CD-ROM drive, data can be read for processing. Now the CD-ROM works like any other drives. Only, it is Read Only Memory!

About Optic Data Storage

The CD-ROM can be compared to a floppy drive, because the disks are removable. It can also be compared with a hard drive, because of similar data storage capacity. Actually, a CD-ROM disk can hold up to 680 MB of data. This equals the capacity of 470 floppy disks. However, the CD ROM is neither a floppy nor a hard disk!

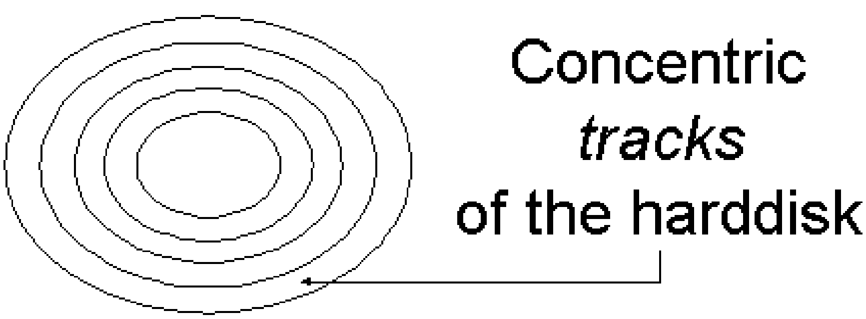
While floppy and hard disks are *magnetic* media, the CD-ROM is an *optic* media. The magnetic media work in principle like an audio cassette tape player. They have a read/write head, which reads or writes magnetic impressions on the disk. The magnetic media contains myriads of microscopic magnets, which can be polarized to represent a zero or numeral one (one bit).

In the optic readable CD-ROM, the data storage consists of millions of indentations burnt into the lacquer-coated, light-reflecting silver surface. The burnt dents reflect less light than the shiny surface. A weak laser beam is sent to the disk through a two-way mirror and the sensor registers the difference in light reflection from the burnt and shiny areas as zero's and one's.

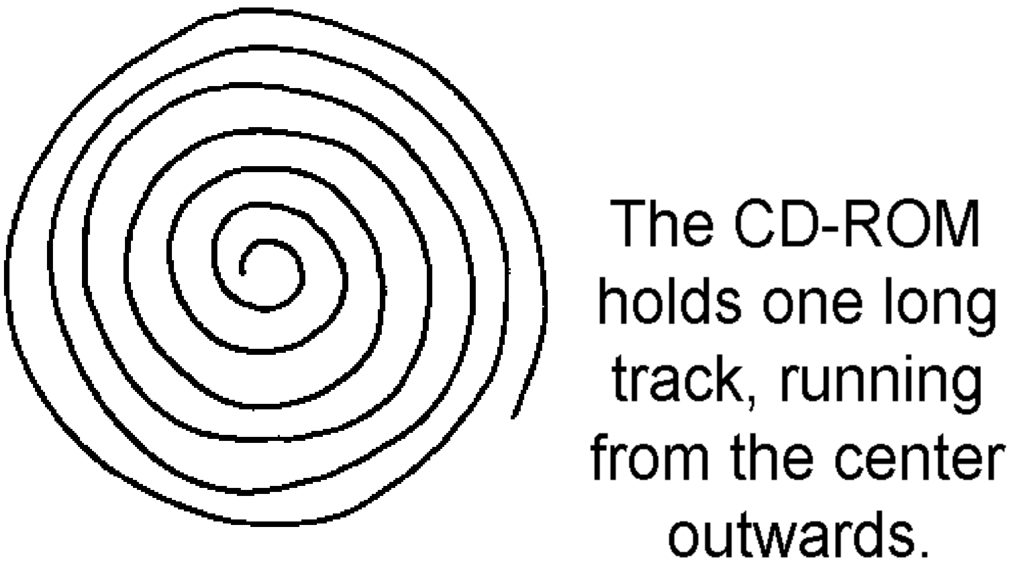
Tracks

Our data consist of bits, each of which is a burnt dent or a shiny spot on the CD-ROM disk. Music CD's are designed much in the same manner. The bits are not splashed across the disk, but arranged in a pattern along the track. Without that organization, you could not read the data.

The platters in hard disks and floppies are organized in concentric *tracks*. There can be hundreds of those from center to periphery:



The CD-ROM is designed differently. It has only one track, a spiral winding its way from the center to the outer edge:



This long spiral track holds up to 650 MB data in about 5.5 billion dots (each is one bit).

Data read from CD-ROM

Data are usually read from the CD-ROM at a constant speed. The principle is called CLV (Constant Linear Velocity). It implies that the data track must pass under the read head at the same rate, whether in inner or outer parts of the track. This is accomplished by varying the disk rotation speed, based on the read head's position. The closer to the center of the disk the faster the rotation speed. In the music CD, data are read sequentially. Therefore, rotation speed variation is not necessary

The CD-ROM disk on the other hand has to read in random pattern. The read head must jump frequently to different parts of the disk. Therefore, it forever has to change rotation speed. You can feel that. It causes pauses in the read function. That is a disadvantage of the CD-ROM media. Also the faster versions can be rather noisy.

Rotation speed and data transmission

There are different generations of CD-ROM drives. Here you see their data.

CD-ROM type	Data transfer speed	Revolutions per minute outermost - innermost track
1X	150 KB/sec	200 - 530
2X	300 KB/sec	400-1060
4X	600 KB/sec	800 - 2120
8X	1.2 MB/sec	1600 - 4240
12X-24X	1.8-3,6 MB/sec	2400 - 6360

The new drives are 24X and 32X spin. When you see their rotation speeds, you wonder how much further this technology can be advanced. The hard disk can spin at high speeds, because it operates in sealed box. The CD does not.

Music from the CD-ROM

The PC CD-ROM drive can play regular music CD's. That is a smart "bonus". It requires three things:

- You must have a sound card in your PC
- The CD-ROM drive must match the MPC-3 multimedia standard (all modern CD-ROM drives do)

- You must connect the CD-ROM drive to the sound card with the short special cable, which comes with the drive.

The CD ROM can easily hold sound data, which can be played directly through the sound card - without use of the short cable I mentioned. It only becomes necessary, when you want to play quality sound music. Certain games (such as *Tuneland*) contain both types of sound.

CD-R and CD-E

In 1990, the CD-ROM technique was advanced to include home burning. You could buy your own *burner*. That is a drive, which can *write* on special CD-ROM disks. These have a temperature sensing layer, which can be changed by writing. You can only write on any given part of these disks once. This CD-R disk is also called a WORM disk (*Write Once Read Many*). Once the CD-R is burnt, it can be read in any CD drive – for sound or data.

There is also a type called CD-erasable (CD-E), where you can write multiple times on the same disk surface. This technique is promising. However, not all CD drives can read these CD's. The latest drives, which can adjust the laser beam to match the current media, are called *multi read.* Look for that, when you buy a new CD-ROM drive.

DVD

The next optic drives we will see in the next few years is the DVD drive. They are being developed by several companies (Philips, Sony, and others) and represent a promising technology. DVD stands for Digital Versatile Disk.

They are thought of as a future all-round disk, which will replace CD-ROM and laser disks. In the future, DVD might also replace VHS tapes for videos.

Certain DVD drives can both read and write the disks.There are also read only, designed for playing videos.

The DVD is a flat disk, the size of a CD - 4.7 inches diameter and .05 inches thick. Data are stored in a small indentation in a spiral track, just like in the CD.

DVD disks are read by a laser beam of shorter wave-length than used by the CD ROM drives. This allows for smaller indentations and increased storage capacity.

The data layer is only half as thick as in the CD-ROM. This opens the possibility to write data in two layers.The outer gold layer is semi transparent, to allow reading of the underlying silver layer. The laser beam is set to two different intensities, strongest for reading the underlying silver layer. Here you see a common type DVD ROM drive:

Other DVD types

We have the following DVD versions:

- **DVD-ROM** is for read-only, like the CD-ROM. This media is usable for distribution of software, but especially for multimedia products, like movies. The outer layers can hold 4.7 GB, the underlying 3.8 GB. The largest version can hold a total of 17 GB.
- **DVD-R** (*recordable*) are write once-only like CD-R. This disk can hold 3.9 GB per side .
- **DVD RAM** can be written and read like a hard disk. Capacity is 2.6 GB per side or whatever the agree on. There are many problems with this format.

Movies and multimedia

A 4.7 GB side can hold 135 minutes top quality video with 6 track stereo. This requires a transmission rate of 4692 bits per second. The17 GB disk holds 200 hours top quality music recording.

The following transmission rates are used in the various media:

- **Music on CD** requires 150 KB per second. That corresponds to the transmission rate in the first 1X CD-ROM drives.
- **Playing of movies** (with a new MPEG-2 compression) requires a transmission rate of about 600 KB per second. That corresponds to the 4X CD ROM drives.
- **Multimedia programmer** is expected to provide1.3 MB per second, like a current 10X CD-ROM drive.

DVD movies are made in two "codes." Region one is USA and Canada, while Europe and Asia is region two. When you play movies, your hardware (MPEG decoder) must match the DVD region. The movies are made in two formats, each with their own coding.

Some DVD drive use Dolby AC-3 sound standard. That is a sound system with five full range speakers to surround you with sound, plus a supplementary low frequency *special effect* channel. To get the full use of the movie sound tracks, you need a AC-3 compatible stereo set up.

The DVD drives come in EIDE and SCSI editions and in 2X, 4X, etc. versions, like the CD-ROM's.

The DVD drives will not replace the magnetic hard disks. The hard disks are being improved as rapidly as DVD, and they definitely offer the fastest seek time and transmission rate (currently 5-10 MB/second). No optic media can keep up with this. But the DVD will undoubtedly gain a place as the successor to the CD ROM.

HD-ROM

Futuristic technology. The well-known optical drive types functions using a laser beam. Norsam Technology has developed an other technology where the controlling beam is made of charged gallium ion particles. This gives a very high density, since the beam is very narrow. The new drive technology is called HD-ROM.

Media	Beam, width	Capacity per disk
CD-ROM	800 nanometer	0,65 GB
DVD	350 nanometer	4,7 GB
HD-ROM	50 nanometer	165 GB

Please read Module 4d about MO drives.

Please read Module 4e about tapestreamers (they are *not* drives).

Please read Module 5c about SCSI.

Please read Module 6a about file systems.

Any CD-ROM drive will work better with the shareware-CD-ROM-cache utility CD-QuickCache. Get it!

To overview

Click & Learn. Module 4d. WWW.MKDATA.DK

MO drives

Magnetic-Optic drives represent an exciting technology. The media is magnetic, yet very different from a hard disk. You can only write to it, when it is heated to about 300 degrees Fahrenheit (The Curie point)

This heating is done with a laser beam. The advantage is that the laser beam can heat a very minute area precisely. In this manner the rather unprecise magnetic head, can write in extremely small spots. Thus, writing is done with a laser guided magnet. The laser beam reads the media. It can detect the polarization of the micro magnets on the media.

MO disks are fast, inexpensive, and extremely stable. They are regarded as almost wear-proof. They can be written over and over again forever, without signs of wear. The data life span is said to be at least 15 years. There are many MO drive variations. The most widespread is the Iomegas Zip drive with the LS120 coming up.

The Zip drive

The Zip drive uses a kind of diskette, which can hold 100 MB. In my opinion, the Zip drive works excellently. They are stable, inexpensive, and easy to work with. The drives are not the fastest. One reason is, that they must make two revolutions for each write operation.

I and many others have used Zip drives since they came on the market. This provides us with a common standard to move large files and to make back-ups. For example, you can use this drive to install Windows 95 on a computer without a CD ROM and avoid having to insert numerous floppy disks.

The 100 MB Zip disk is borderline size. However, compared to the work I had to do previously, compressing files with PKZIP onto multiple diskettes, these are very practical.

Two types of interface

The Zip drive exists in different versions:

- **Internal and external**
- **For SCSI and to floppy/parallel port**

The SCSI model is by far the fastest. That is really good. If your SCSI controller is installed with Windows 95, you just have to install the drive with two screws and two cables and you are in business.

The parallel port version is good, because it can be connected to *any* PC. I have a boot diskette, which includes a driver plus the program GUEST.EXE. I connect the drive to a parallel port, and boot with the diskette. Then it is ready to run.

I have the quite fast SCSI version installed in my stationary PC. I use the somewhat slower parallel port version "in the field."

My latest information is that 5 million Zip drives have been sold. This just about makes it a de facto standard. The BIOS manufacturers AMI and Phoenix include the floppy version of the drive in their programs as a boot device. That will eliminate the need for other drivers, and you will be able to boot from the Zip-disk.

Other MO drives

There are other MO drives available. However, they are currently in a quite different price range than the Zip drives.

In the last four years, we have heard about the LS120 drive and now it is finally available. It is a Compaq 120 MB standard. It is supposed to replace the regular floppy drives. At the same time, they read floppy diskettes much faster than the traditional floppy drives.

The LS120 ought to become the new floppy standard, but it has come to late. Soon we will have the DVD RAM disks which holds from 2.6 GB and up.

Sony has a MO drive called HiFD holding 200 MB on a 3½" floppy disk. It also reads DD and HD floppies.

Use the LS120 or the HiFD in new PC's - they are cheap and good for backup. The drives use EIDE interface.

Module 4e about tape streamers (they are *not* drives).

Module 5c about the SCSI-interface

Module 6a about the file systems.

Click & Learn. Module 4e. WWW.MKDATA.DK

About tape streamers

Tape streamers are used for data *backup*. They come in different types and price ranges, but have this in common:

- The tape streamer does not work like a *drive*. You can not retrieve any particular file. The data must be read using special back-up software.
- Data are stored *sequentially* on the tape. This means that you can not, contrary to disks or CD-ROM's, read in random fashion. You must wind the tape to the desired location.

The advantage of Tape streamers is their low cost. They contain lots of data on inexpensive tapes. They are available in different types:

Tapestreamers - variations	
Installation	• *Internal* in the PC • *External* in a box connected to the PC
Interface	• On floppy controller or parallel port • EIDE (rather new) • SCSI
Tape types	• QIC 40/80/Wide • Travan TR1/TR3/TR4 • DAT

Tapes		
Standard	Capacity	Typical interface
QIC	120 MB	Floppy/parallel
Travan TR1: TR3: TR4:	 800 MB 1,6 GB 4 GB	 Floppy/parallel Floppy/parallel EIDE or SCSI
DAT	2/4 GB	SCSI

To overview

Click & Learn. Module 5. WWW.MKDATA.DK

About I/O units, expansion cards, adapters, etc.

Intro to I/O

This page should preferably be read together with module 2c, 2d, 5b and 5c. The first two describe the I/O buses and the chip sets. Here we will look at the other end of the I/O buses, the "exit."

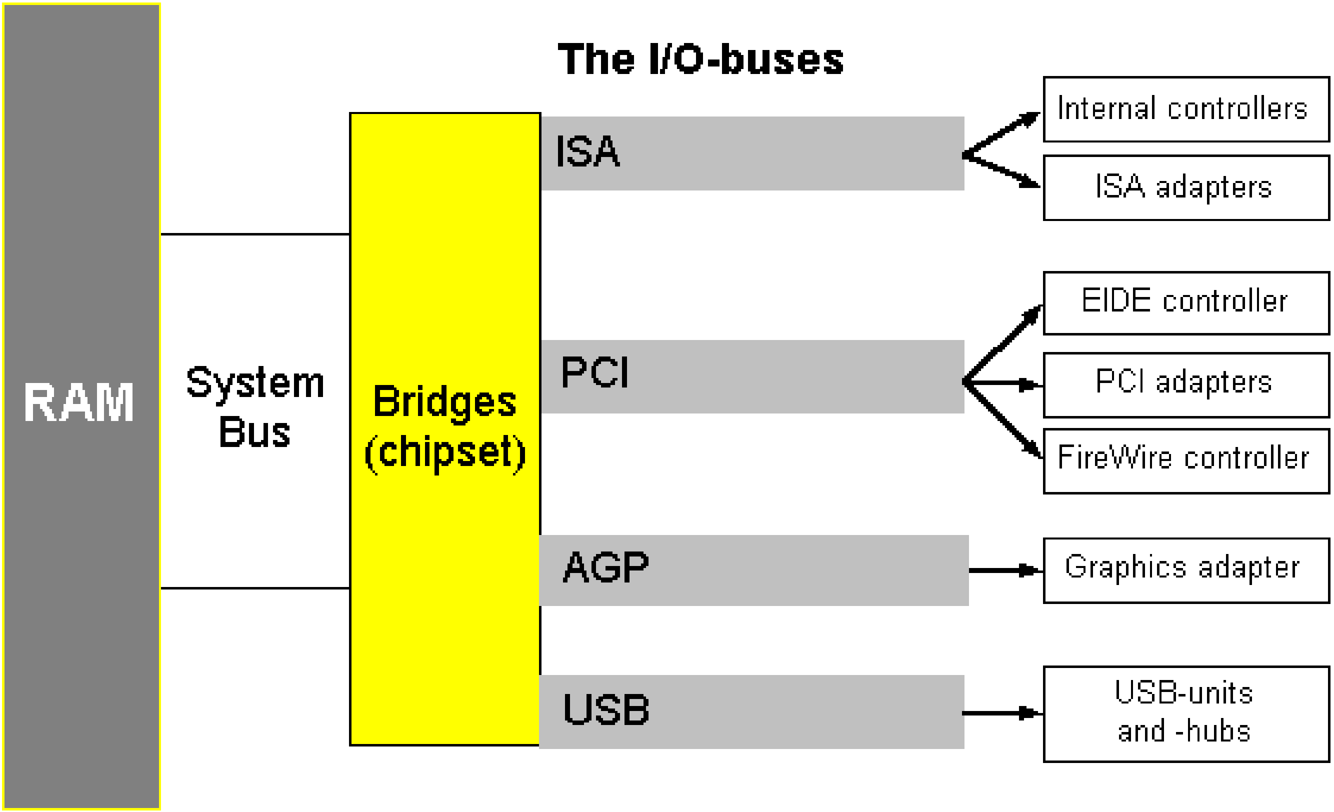
There are four I/O buses in the modern PC architecture and each of them has several functions. They may lead internal and external ports or they lead to other controlling buses. The four buses are:

- ISA, which is old and hopefully soon disappears.
- PCI, which is the newer high speed multifunction I/O bus.
- AGP, which only is used for graphics adapter.
- USB, which is the new low-speed I/O bus to replace ISA.

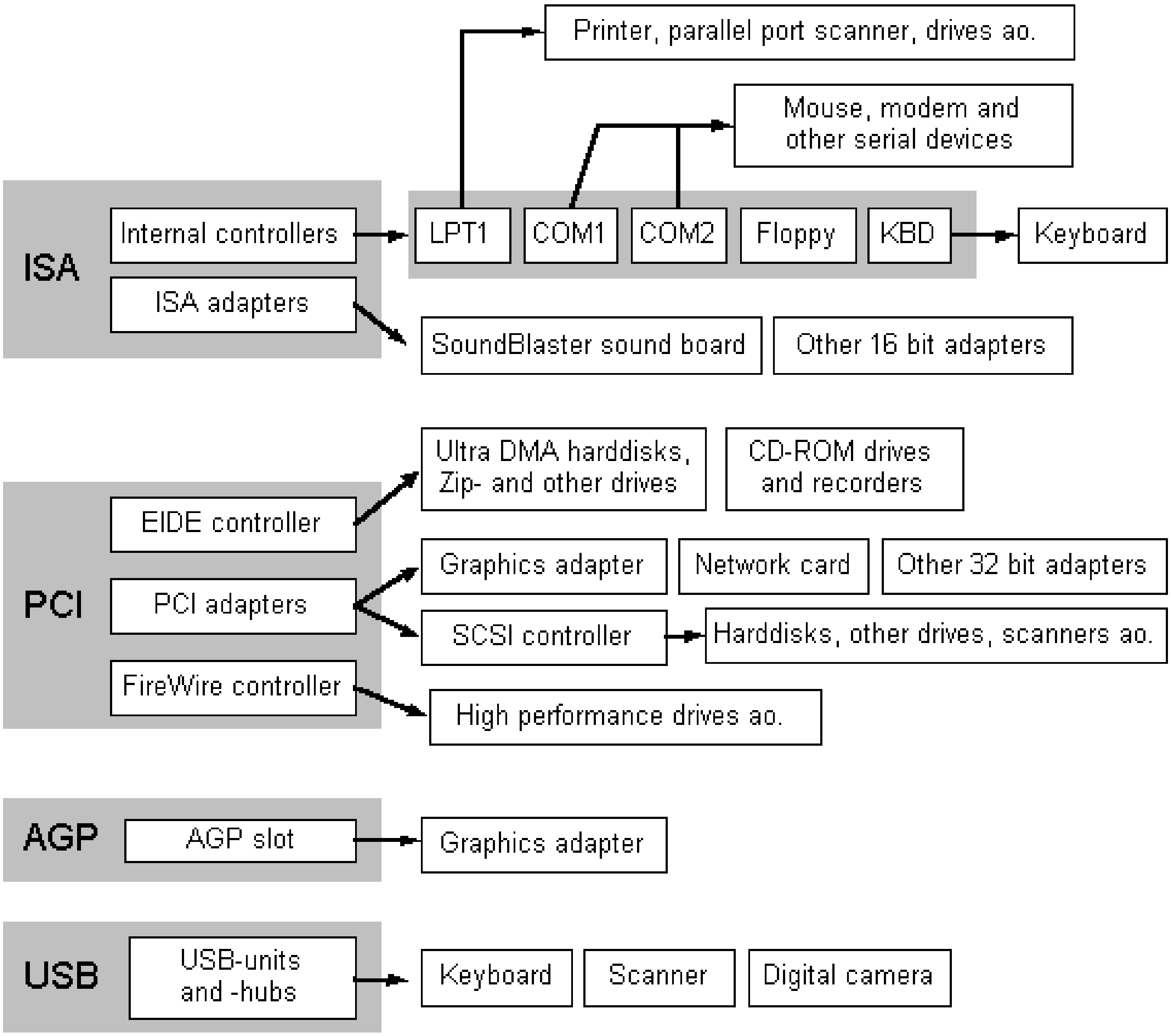
The ISA and the PCI bus both end in a twofold exit:

- Internal I/O ports (LPT, KBD, COM1, COM2, EIDE ao.)
- Expansion slots in the system board, in which we can insert adapters.

If you look at this illustration you will se the overview of this architecture:



If we focus on the right end of the illustration we approach the I/O units. Here you get a closer look at that:



As you see, there is room for a lot of units to be connected to the PC.

The PCI bus is the most loaded of all the buses. It is used for so many purposes that the output for the graphics adapter has been isolated on its own AGP-bus. But still the PCI bus is under heavy load, connecting the system bus to the network controller and the various EIDE- and SCSI-drives. Hopefully the FireWire bus will become separated from the PCI bus in future architecture.

The internal I/O ports

As mentioned, the USB is going to become the main bus for low-speed devices. But so far we still use the internal "face" of the ISA bus for a range of purposes. At any PC motherboard you find these:

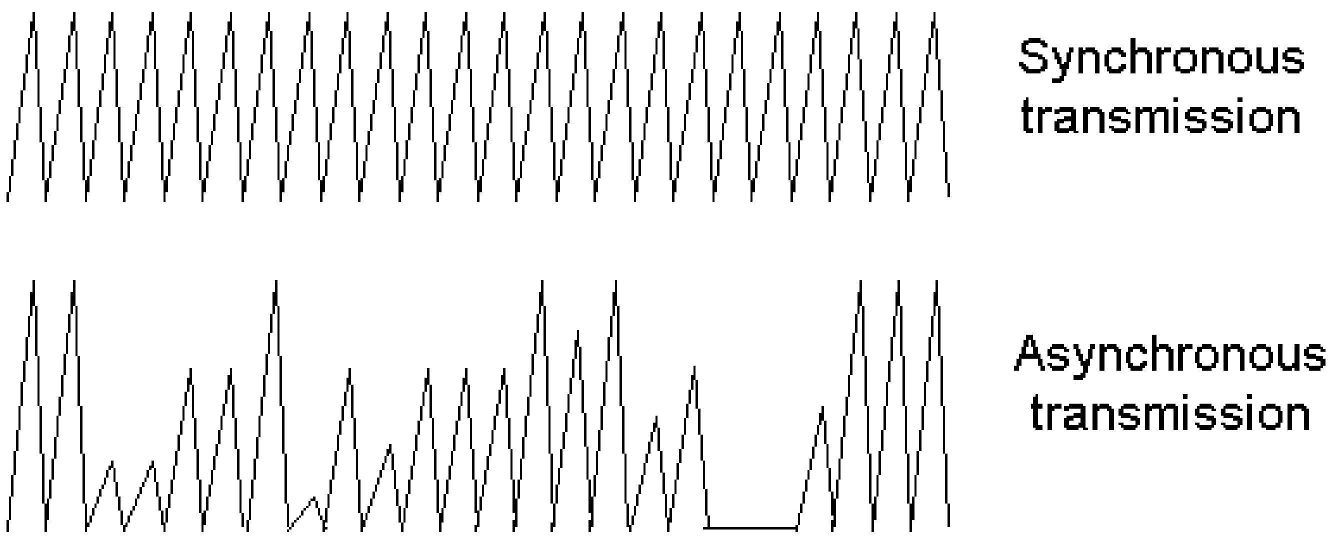
- The floppy controller
- The serial ports

- The parallel port(s)
- The keyboard controller

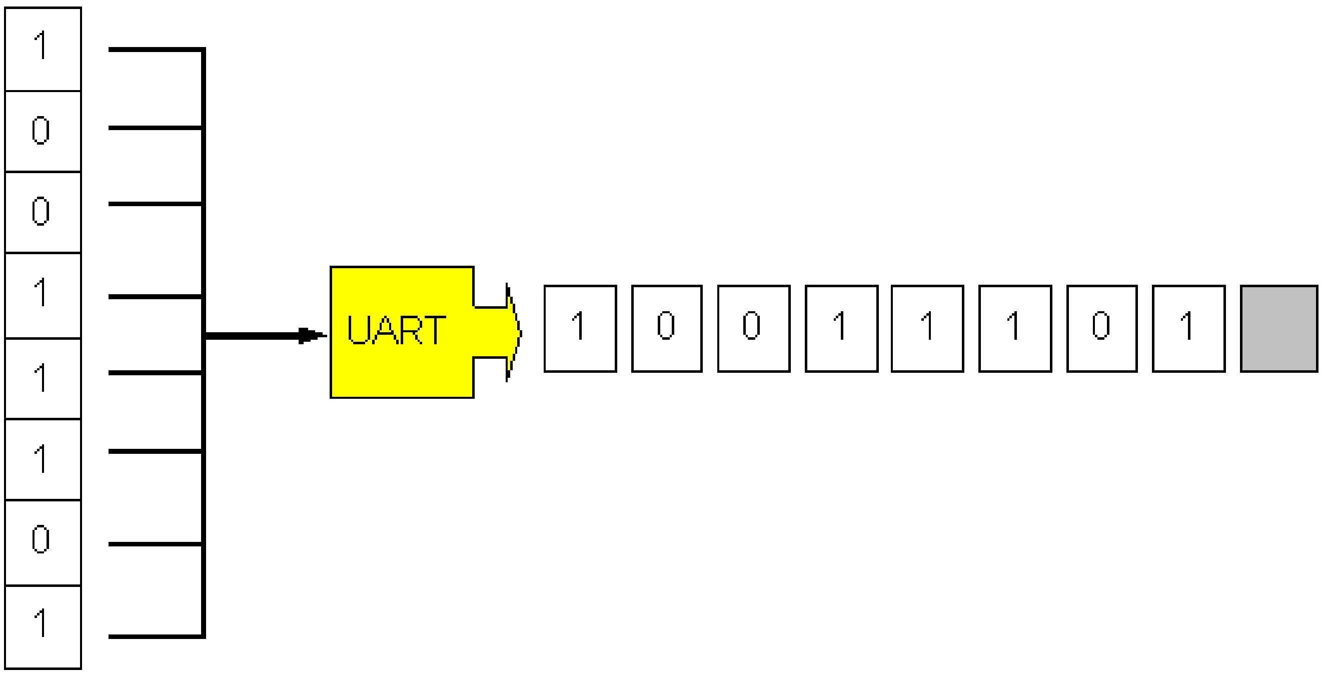
They all occupy IRQ's which is a central part of ISA architecture and a pain in the a... Let us take a moment to look at these ports and controllers.

The serial ports

Serial transmission means to send data from one unit to another one bit at the time. The PC architecture traditionally holds to RS232 serial ports. The RS-232 standard describes an asynchronous interface. This means that data only are transmitted when the receiving unit is ready to receive them:



The serial ports are controlled by an UART chip (*Universal Asynchronous Receiver Transmitter)* like 16550 AFN. This chip receives bytes from the system bus and chops them up into bits. The most common package is called 8/N/1 meaning that we send 8 bits, no parity bit and finally one stop bit. This way one byte occupies 9 bits:



The serial transfer is limited to a speed of 115,200 bits per second. The cable can be up to 200 meter long. The serial ports can be used to connect:

- The mouse
- Modems
- ISDN adapters
- Printers with serial interface
- Digital cameras
- ....

These units are connected to the serial ports using either DB9 or DB15 plugs.
In a few years time all these devices will connect to the USB bus instead.

The parallel port

Parallel transmission means that data are conducted through 8 separate wires - transmitting a full byte in one operation. This way the parallel transmission is speedier than the serial, but the cabling is limited to 5-10 meters. The cable is fat and unhandy, holding up to 25 wires and the transmission is controlled according to the Centronics standard.

Most printer manufactories use a 36-pins Amphenol plug, where the PC's parallel port holds a 25-pinned connector. Hence the special printer cable. To the left you se the 25 pin connector, to the right the 36-pin:

Then parallel port represents the most uncomplicated interface of the PC. It is always used to connect the printer, but with the bi-directional parallel port (EPP/ECP), other devices have found their way to this interface. Today you find:

- ZIP-drives
- Portable CD-ROM-drives
- SCSI adapters
- Digital cameras
- Scanners

all using the parallel port to connect to the system bus.

The EPP/ECP ports

Today we operate with Enhanced Parallel Port/Enhanced Capability Ports. This method for bi-directional (half-duplex) parallel communication offers higher rates of data transfer (up to 1 megabyte per second) than the original parallel signaling method. EPP is used for for non-printer peripherals, where ECP is for printers and scanners. You find the settings for the printer port in the setup program on the motherboard.

Both port types are parts of the IEEE 1284 standard, which also includes Centronics.

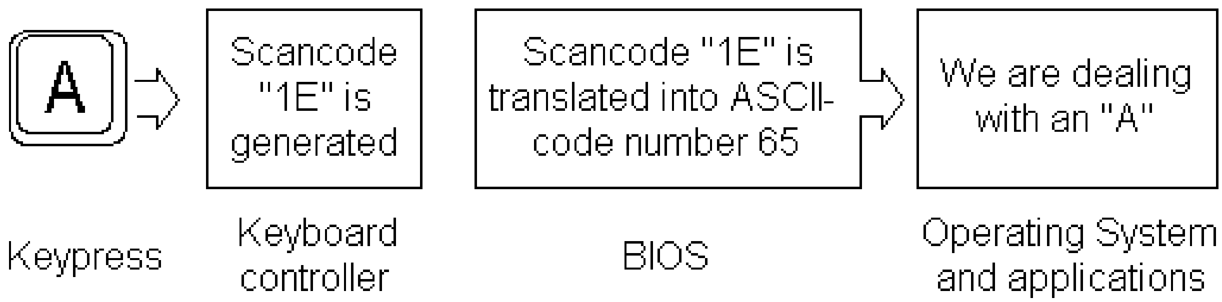
To get the best results all the involved hardware and the operating system has to be EPP/ECP compatible. Windows 95 supports IEEE 1284 in its parallel plug-and-play feature. It also supports ECP if you have a printer and a parallel port with ECP. The printer cable has to be complete with all 25 wires connected.

The keyboard

Traditionally the keyboard is connected using a DIN or PS/2 mini-DIN plug. Soon we shall have USB keyboards but the old ones connect to the internal ISA bus occupying an IRQ.

The keyboard operates with scan codes, which are generated each time a key is pressed and released.

The scan codes are translated into ASCII-values, which are translated according to the code pages (se module 1a and 1b). Here you see a simple illustration of the system:



This system is quite smart since it allows all kind of programming of the keyboard. Each key generates a unique scan code. This happens completely independent of the typeface that is printed on the plastic key.

At the other end, the code pages represents a programmable interpretation of the key press; you can assign any type to any key as you want it. Languages like German and French use different keyboard layouts as well as many other languages.

Adapters

In a stationary PC, adapters are typically printed circuit boards called expansion boards or expansion cards. They form a link between the central PC unit and various peripherals. This is the so-called open architecture.

- Typically, adapters provide functions, which are *separated* from the system board.
- Adapters provide *expansion* capability to the PC.

There are PC's without expansion slots. In that case all functions must be built into the system board. You could easily include chips for graphics, ethernet, SCSI, and sound on the system board. This is not common in stationary PC's. Portable, *laptop PC's* have nearly all electronics on the system board. This is called closed architecture.

A traditional PC has a system board which contains all standard functions (except the graphics chip). To this system board you can add various expansion cards, which *control* one or more peripheral units:

The system board	Expansion boards
Standard functions incl. control of keyboard, COM and LPT ports. and four EIDE units.	Video card Network controller Sound card SCSI card 3D graphics controller (for 3D games)

Other expansion board types:

- **Internal modem (in lieu of external modem)**
- **ISDN adapters**

- **Extra parallel ports**
- **Video editing boards**
- **Special graphics cards, which supplement the usual (3D and MPEG)**
- **TV and radio receivers.**

The integrated hard disk controller

In the Pentium based PC, the hard disk is connected to an EIDE controller, which is integrated on the system board. Likewise, the serial and parallel ports are connected directly to the system board. This is new. On the 386 PC's, you had to install a special controller cards (I/O cards) to handle these functions. They are included in the modern chip sets on the system board. Other functions are not integrated. That includes:

The video controller

You have to install a video card to make the PC function. It would be illogical to assemble a PC without a video card. You would not be able to see what you are doing, since the video card governs data transmission to the monitor.

The advantage of this design is, that the user can choose between numerous video cards in various qualities. A discount store may offer a *complete Pentium based PC* (without printer) and with the cheapest video card for $669.-. If the buyer is quality oriented, he would want to spend an additional $40 to get a much better video card.

The modular PC design

In this way, various expansion boards provide flexibility in assembling a customized PC. At the same time, various electronics manufacturers are specializing their production:

ASUS and Tyan are good at making system boards. Others, like S3, Matrox, and ATI specialize in making graphics chips and expansion boards. Olicom make only net boards. Adaptec make only SCSI controllers and Creative Labs make SoundBlaster sound boards.

This variety of manufacturers offers the consumer wide choices. Your PC can be customized and configured according to your needs and wallet size.

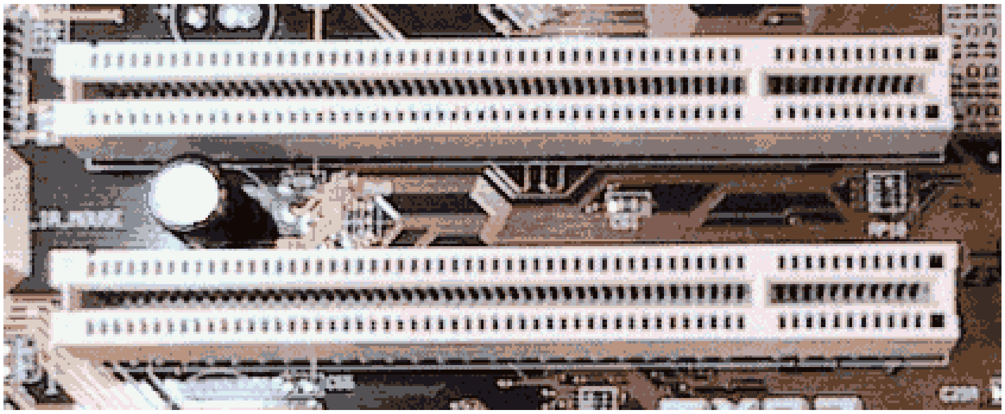
About the electronics

The adapter is a printed circuit board. They have an *edge connector*, so they can be inserted in expansion slots in the system board. The expansion slots connect to the I/O busses. Since the Pentium system board has two I/O busses, it has two types of expansion slots:

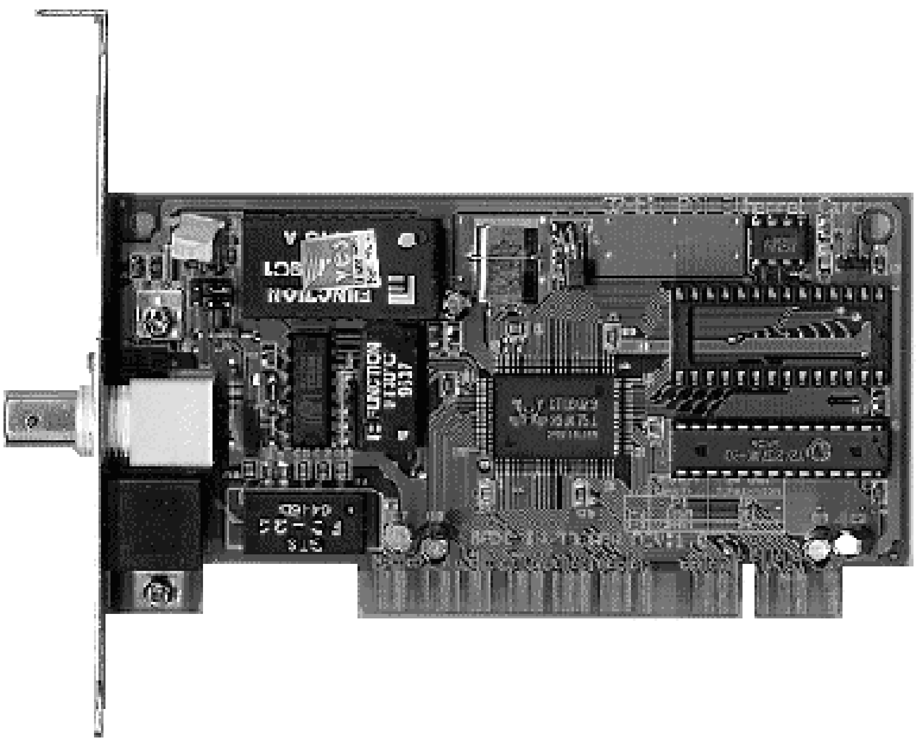
- **ISA slots**
- **PCI slots**

Typically, on a regular Pentium system board there are three or four of each type. That gives a total of 7 expansion slots. One expansion board can be installed in each of these. You simply press the edge connector of the expansion board into the expansion slot. Now it is connected to the bus.

Here you see two PCI slots open for video cards, network controllers and others:



Below, you see a *network adapter.* It is an ethernet card with PCI interface, so it fits in a PCI slot in the Pentium. This inexpensive board allows your computer to join a network with other net board equipped PC's. Please compare the edge connector at the button of the card with the sockets above. They fit together!



IRQ's

When you install an expansion board in a slot, it gets connected to the I/O bus. Now the board can send and receive data. But who regulates the traffic? Who gives clearance to the new controller to send data? It would appear that data traffic could soon be chaotic.

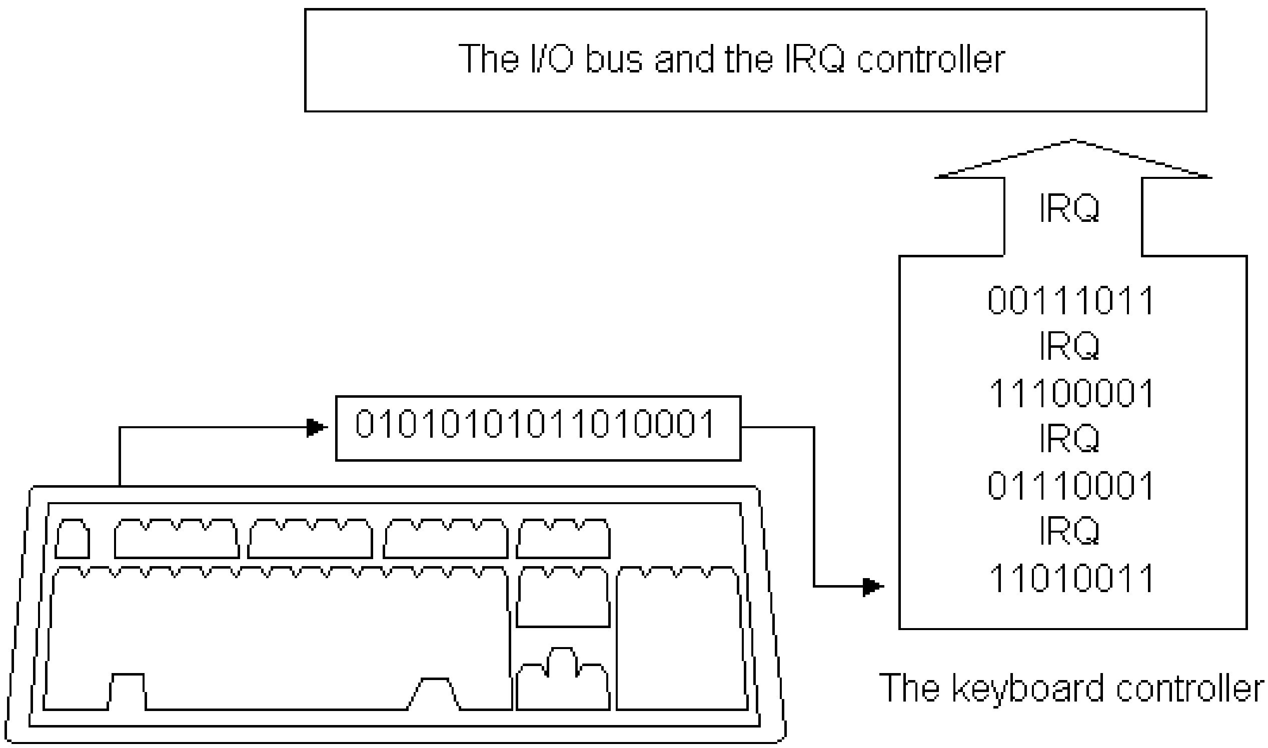
To control data traffic on the I/O bus, the concept of IRQ (*Interrupt ReQuest)* was created. Interrupts are a fundamental principle in the PC design. There are two types of interrupts: Software Interrupts are used to call any number of BIOS routines. Hardware Interrupts are the subject of this page.

Hardware Interrupts

The adapter or unit on the I/O bus uses the interrupt to signal request to send or receive data. An interrupt signal is like a door bell. The unit signals by applying a voltage to one of the wires in the bus - an IRQ. When the CPU acknowledges the signal, it knows that the unit wants send or receive data, or is finished.

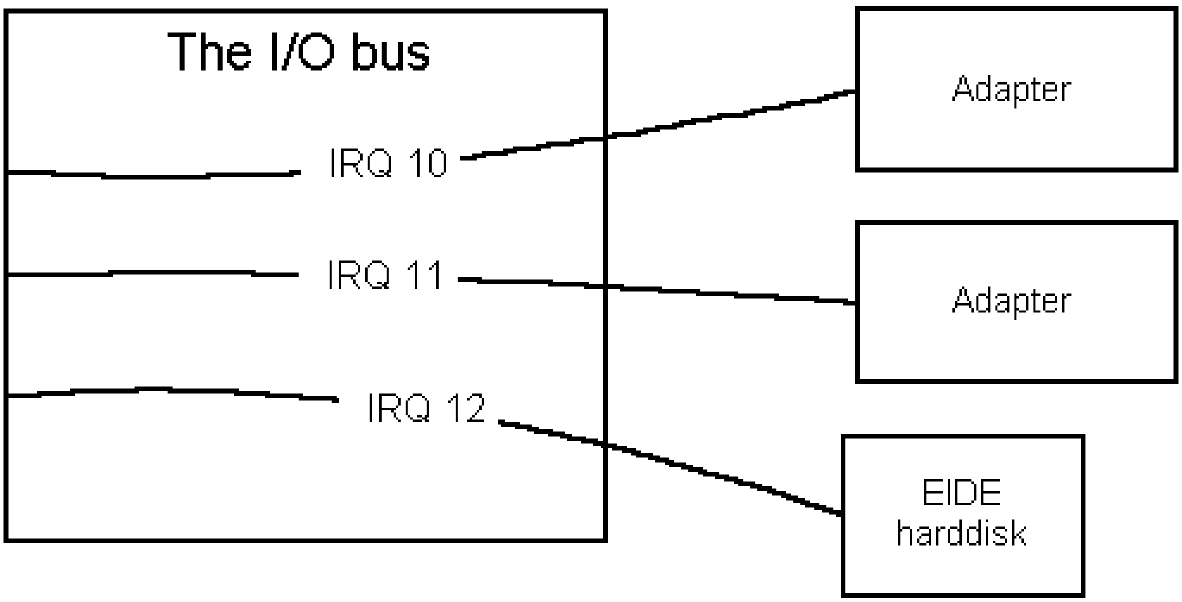
The advantage of IRQ's is that the CPU can manage other tasks, while an adapter "massages" its data. When the adapter has finished its task, it will report to the CPU with a new IRQ.

As an example, let us see how keyboard data are handled. The keyboard send bits, serially, through the cable to the keyboard controller. The controller organizes them in groups of 8 (one byte). Every time it has a byte, it sends an IRQ to the I/O bus. The IRQ controller asks the CPU permission to use the bus, to send the byte to wherever. The IRQ controller reports back to the keyboard controller, giving clearance to send the next character (byte):



IRQ wires

Physically, the IRQ is a wire on the bus. This wire connects to all expansion slots. Therefore, regardless of in which slot you install an adapter, the adapter can communicate with an IRQ. The PC is "born" with 15 IRQ's, but five of them are internal, and can not be used with I/O cards. We find 10 accessible IRQ's on the I/O busses. Each of those consist of a circuit board wire, which goes through the entire bus. When you install an expansion card in a vacant slot, one of the IRQ's is assigned to it.



When a signal arrives on an IRQ channel, that is a message to the CPU. It is told that a unit wants to get on the bus. *Which* unit is to be identified through the *IRQ number*.

Next the unit is admitted to the bus, to send or receive data. When the transaction is completed, another signal is transmitted to the CPU to indicate that the bus is vacant.

The IRQ's have different *priorities*, so the CPU knows which IRQ have priority, if two signals are sent simultaneously.

The IRQ system is guided by a controller chip, like Intel 8259. It can handle 8 IRQ signals and couple two of them together, via IRQ 2 or 9. All PC's with ISA bus include two 8259 chips.

MSD (Microsoft Diagnose System)

Let me show an image of the MSD diagnostic program, which you can run in Windows 95. It shows the use of IRQ's on a PC:

IRQ	Address	Description	Detected	Handled By
0	CC00:0000	Timer Click	Yes	Unknown
1	0A2C:08D2	Keyboard	Yes	KEYB
2	F000:FF47	Second 8259A	Yes	BIOS
3	F000:FF47	COM2: COM4:	COM2:	BIOS
4	F000:FF47	COM1: COM3:	COM1:	BIOS
5	F000:FF47	LPT2:	No	BIOS
6	0956:009A	Floppy Disk	Yes	Default Handlers
7	0070:0465	LPT1:	Yes	System Area
8	0956:0035	Real-Time Clock	Yes	Default Handlers
9	F000:F7C9	Redirected IRQ2	Yes	BIOS
10	F000:FF47	(Reserved)		BIOS
11	F000:FF47	(Reserved)		BIOS
12	0956:00E2	(Reserved)	Not Detected	Default Handlers
13	F000:2930	Math Coprocessor	Yes	BIOS
14	0956:00FA	Fixed Disk	Yes	Default Handlers
15	F000:FF47	(Reserved)		BIOS

MSD shows the IRQ's of the PC, where the program is run. There are a total of 15 IRQ channels and each IRQ is assigned to a unit. However, it is not always possible to utilize IRQ 9. It functions like a bridge between two parts in the IRQ system.

In the above illustration, IRQ numbers 5, 10, 11, 12, and 15 appear vacant.

IRQ numbers 2 and 9 show the linking between those two IRQ controllers.

Some IRQ's are reserved for various internal units, which must also be able to disconnect the CPU. Those are IRQ numbers 0, 1, 2, 8, and 13, as you can see in the illustration above. They are not available for other units. In principle, the remainder are available for expansion boards and EIDE units.

IRQ's are assigned during the PC start-up. An ISA expansion board is assigned a given IRQ during start-up. That IRQ is used every time that expansion board uses the bus.

Shared IRQ's

The modern I/O busses MCA, EISA and PCI permit shared IRQ's. Thus, two adapters can share one IRQ. When the IRQ is activated, the drive programs for the two adapters are checked, to identify which is on the bus.

IRQ and conflicts on the ISA bus

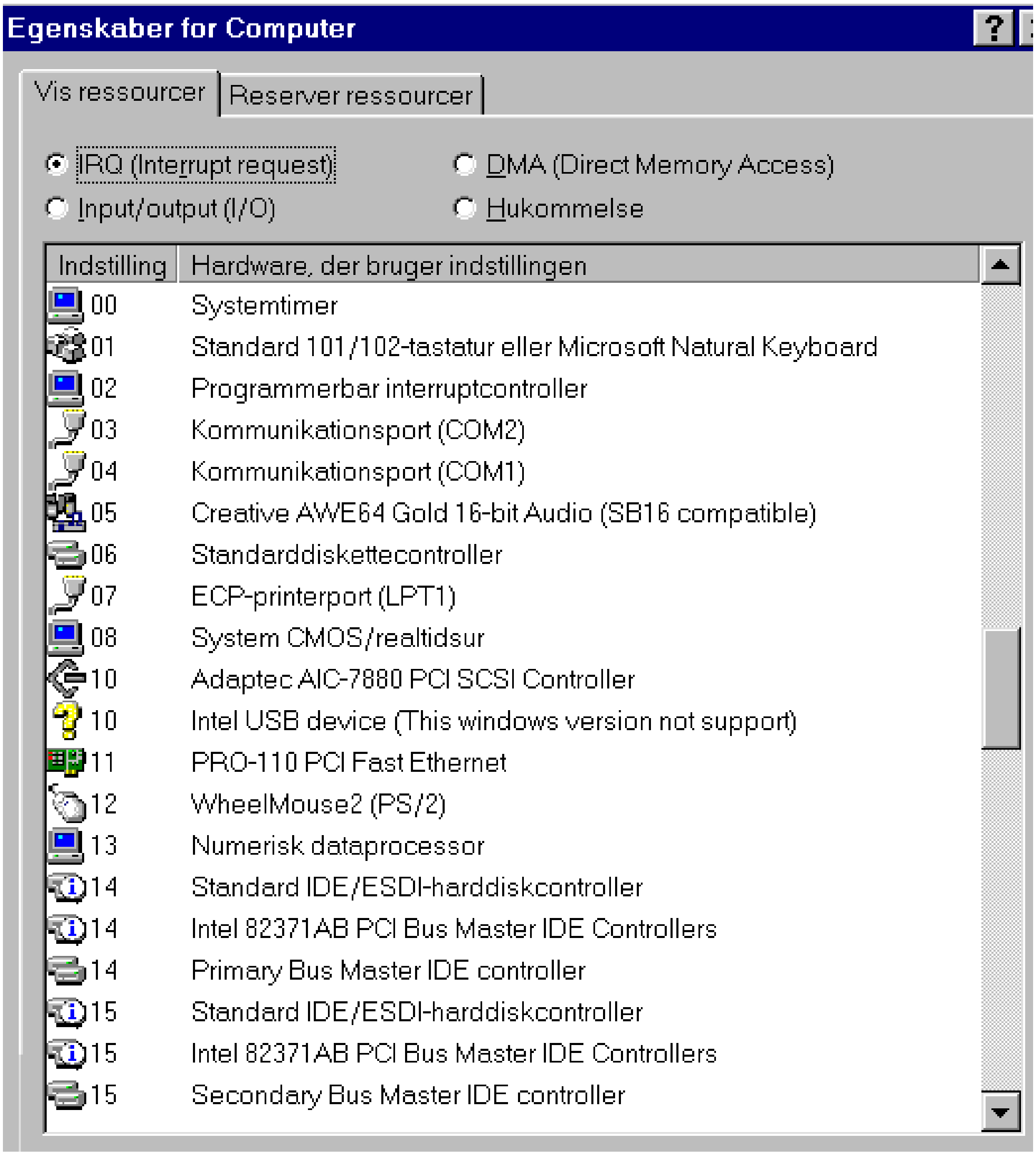
The IRQ system can cause some problems on the unintelligent ISA bus. When bus and adapters are referred to as unintelligent, it implies that they are unable to organize the IRQ distribution on their own.

In order to function, an ISA network controller must be assigned an IRQ. The manufacturer could preset it to work with IRQ 9, 10, 11, or 12. One of these values, let us say IRQ 11, is preset as the *default* value. When the customer installs that board, during start-up it will try to access the bus as IRQ 11. If no other units is connected to IRQ 11, it should work. If IRQ 11 is occupied, we have a problem. Those two units would get in a *conflict*. Often, the PC will not start at all and panic erupts.

The solution is to change the IRQ of the adapter. The manufacturer has designed the board to work on IRQ 9, 10, 11, or 12. Number 11 was the default. If that does not work, you must adjust to another. This can be done with the accompanying software, or by resetting a little *jumper* - an electric contact on the board, which has to be reset. The manual for the board will include instructions about how to do this.

These IRQ problems can be a terrible nuisance. If both sound and net boards had to be installed in ISA slots in the same PC, sometimes I had to give up.

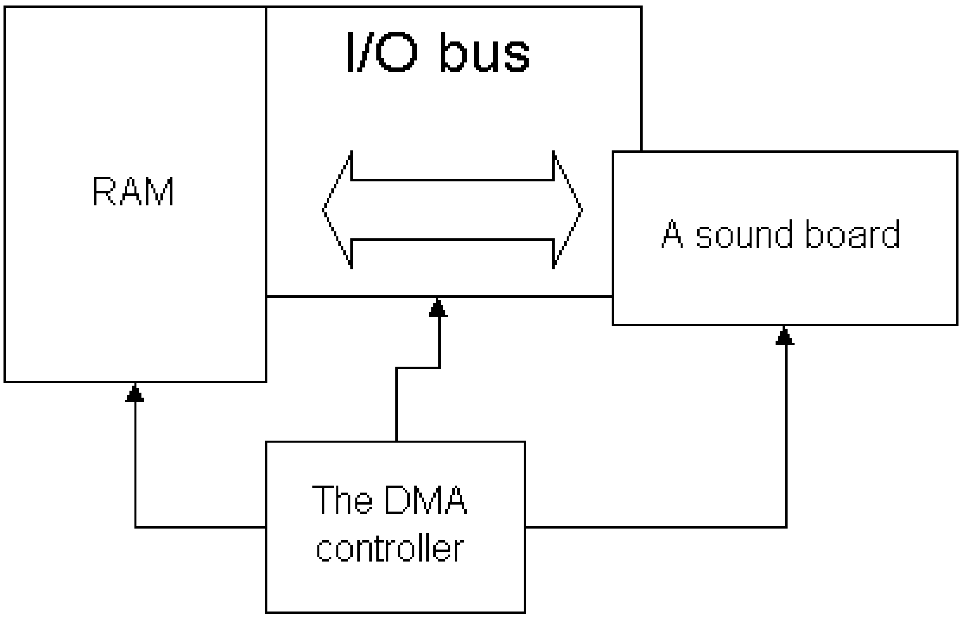
In Windows 95 (System, Computer, Properties) you can find an excellent overview of the IRQ's. Here it is from my Danish version:



DMA

IRQ's are only one of the problems with ISA boards. The other one is DMA (Direct Memory Access). That is a system which allows an adapter to transfer data to RAM without CPU involvement.

Normally, the CPU controls all bus activities. With DMA, this "intelligence" is assigned to a DMA controller on the system board. This special controller chip (Intel 8237) has clearance to move data to and from RAM, via the I/O bus, without burdening the CPU.

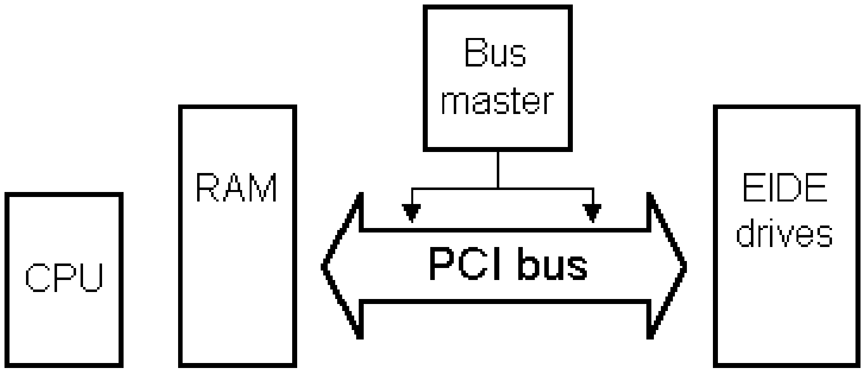


You can implement a number of DMA channels, which can be used by the ISA boards. Each channel has its own number and one controller can be in charge of four channels. Each ISA unit can occupy one of these channels, if so designed. Diskette drives utilize DMA.

The DMA system can result in conflicts between two units on the bus, if they have requested the same DMA channel. As an example, on ISA sound boards you have to reset both IRQ and DMA number.

Bus mastering

There are no DMA channels on the PCI bus. In stead *bus mastering* is employed. It is a similar system, where special controller functions allow adapters to control the bus. Thus, they can deliver their data directly to RAM, minimizing the workload on the CPU. It does not need to keep track of the transactions, the bus master takes care of that.



This allows the PC to multitask, handle more than one task at a time. The hard disk can pour streams of data to RAM, while the CPU handles some other task. The bus mastering system works *fairly well* with EIDE hard disks. However in this particular area, the SCSI controller is far more advanced. EIDE bus mastering is rather new and we will see further developments in this area.

Bus mastering version	Chip set	Year
DMA mode 2	82430FX	1995
Ultra DMA	82430TX	1997

The latest version Ultra DMA is described in module 5b.

I/O addresses

Finally, we need to mention how the the CPU finds all these units - adapters, ports. etc. They all have an address - an I/O port number.

Each unit can be reached through one of more I/O ports. Each port is a byte port. That means that 8 bits (one byte) can be transmitted simultaneously - parallel mode.

If the unit is on the ISA bus, it handles 16 bits at a time (*words*). Then you link two consecutive ports together, to make a 16 bit channel. If we talk about about a 32 bit PCI unit, we link four byte ports together to get 32 bits width (called *dword).*

The PC has a built in listing of all I/O units, each of which has their own "zip code" - a port address. Since the PC is basically a 16 bit computer, there are 2 in the 16 power possible addresses (65,536) - from 0000H to FFFFH. They are described in the hexadecimal number system as 5 digit numbers. Hexadecimal is a 16 digit number system. Digits go from 0 to 9 and continue with 6 letters A - H. Let me show you some examples of I/O addresses:

Unit	I/O ports
CMOS RAM	0070H

Keyboard	0060H ... 0063H
Serial port 1 (COM 1)	03F8H ... 03FFH
Parallel port 1 (LPT1)	0378H ... 037FH

Fortunately, you do not have to adjust port addresses too often. Some adapters give room to adjust to user option I/O addresses, but you have to have bad luck to encounter any conflict in this area.

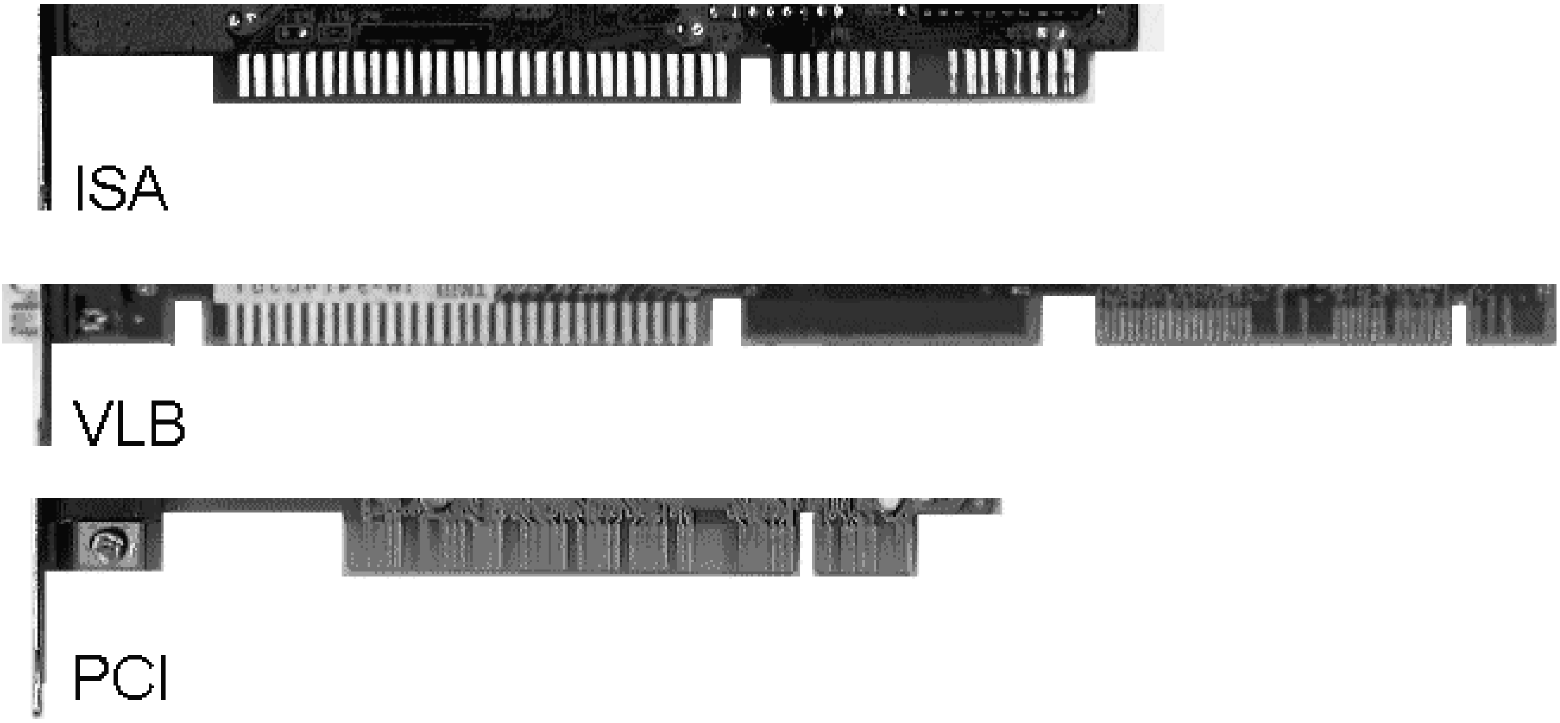
Plug and Play

Plug and play (PnP) is an industry standard for expansion boards. If the board conforms to the PnP standard, the installation is very simple. The board configures itself automatically. These are the minimum requirements:

- The PC system board must be PnP compatible.
- The operating system must be capable of utilizing PnP. Currently, only Windows 95 can do that.
- The adapter must be able to inform the I/O bus which I/O addresses and IRQ's it can communicate with.
- The adapter must be able to adjust to use the I/O address and the IRQ, which the I/O bus communicates to the adapter.

A look at the adapter edge connector

The different I/O cards each fit with a particular I/O bus. The different busses each have their own system board slot configuration. That is a socket in the system board, in which you press in the expansion board. Here you see three different edge connectors fitting into each their type of socket. The ISA bus has a total of 98 prongs (31+18 on each side).



PC Card

In portable PC's, the adapter is usually placed in a PCMCIA slot (a so-called PC card). The first generation PC cards were technically connected to the ISA bus. The newer ones (PC Card to Card Bus) are connected to the PCI bus. Here you see a network controller, as a PC Card. It is about the size of a credit card, but slightly thicker:

The PC Card is placed in a special socket, where it can be inserted or removed, while the PC is operating. Actually, each socket acts like an I/O unit, regardless of whether there is a PC Card in it or not. When the card is inserted, it is automatically configured with I/O address, IRQ, etc. Windows 95 provides by far the best support for PC Cards.

Two function adapters

Integrated adapters with more than one functions are space savers. Especially, the ASUS company has introduced dual function boards to stationary PC's, since they utilize both the ISA and PCI bus to share a slot:

- **Graphics + sound**
- **SCSI + sound**

There are also two function PC Cards for portable PC's:

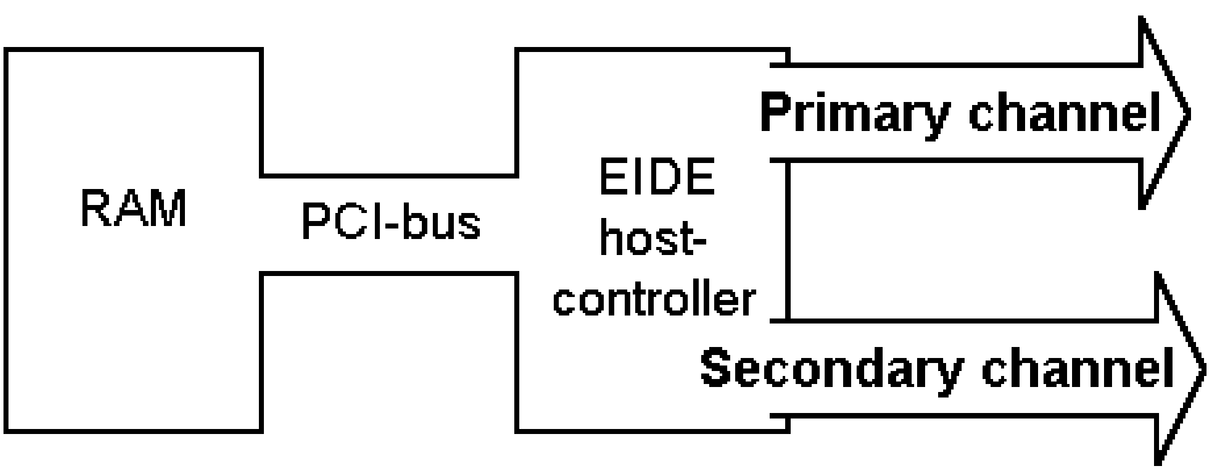
- **Ethernet network controller + modem**
- **Token Ring network controller + modem**

To overview

About Interfaces. EIDE, Ultra DMA and AGP

What is EIDE?

EIDE is the current standard for inexpensive, high performance hard disks. EIDE stands for Enhanced IDE and that is precisely what it is. Some manufacturers (like Seagate) call it ATA. You can think of EIDE as a *bus* - which is a *host controller* - which controls it, and you can connect up to four units. Here you see the controller and its two channels:



All Pentium system boards since 1995 have this EIDE controller built into the chip set. That allows the hard disk and other EIDE units to be connected directly to the system board.

Improvements

The EIDE standard is a great improvement over the old IDE. Here are some examples:

- **The Hard disk** can exceed the 528 MB IDE limit. Currently the largest EIDE disks are 8.5

GB and this number keeps increasing.

- **The hard disk's interface** is moved from the ISA bus to the high speed PCI bus.
- **Four units** can be connected to the system board, which has two EIDE channels. Each channel can be connected to a *master* and a *slave* unit.

The most important feature is the interface directly on the PCI bus. This has given EIDE transfer speeds and disk capacities, which far exceed older controller principles. Concurrently, there is a continual development of the *protocols,* which are needed for the connection between the units and the EIDE bus.

Four units

The EIDE interface is not designed for hard disks only. There are four channels, which can be connected to four independent units:

- Hard disks (which must be on the primary channel. On some system boards, this has the greatest transfer capacity)
- CD ROM drives
- DVD drives
- LS-120 drives and other MO drives
- Tape streamers

EIDE is thus designed as an inexpensive all-around interface, which can be connected to all kinds of storage media.

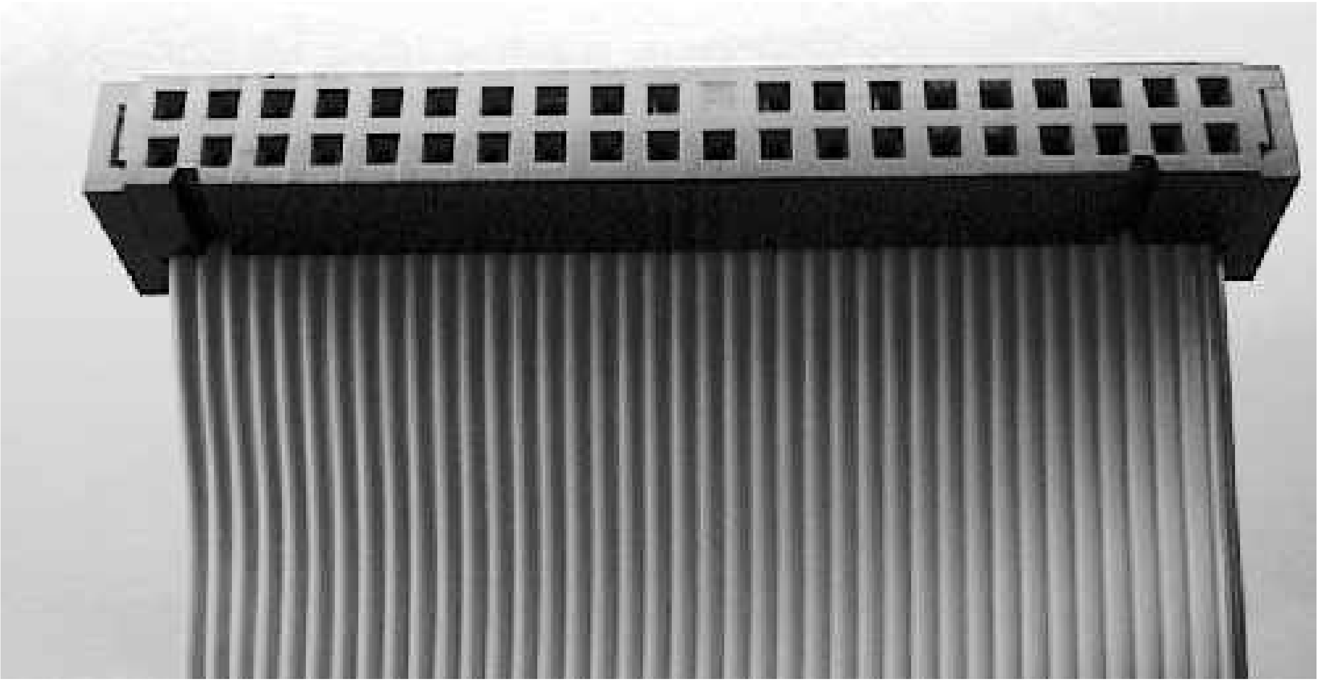
Auto detect

The BIOS on the system board has a neat auto detect feature, which often allows EIDE units to be connected directly and work immediately. The PC start up program automatically finds the necessary information about the drive via the auto detect function.

Sometimes you have to assist the hard disk installation by activating the auto detect in the CMOS Setup program, but often it runs by itself. You definitely do not have to key in information about cylinders, etc., as you had to with earlier IDE units.

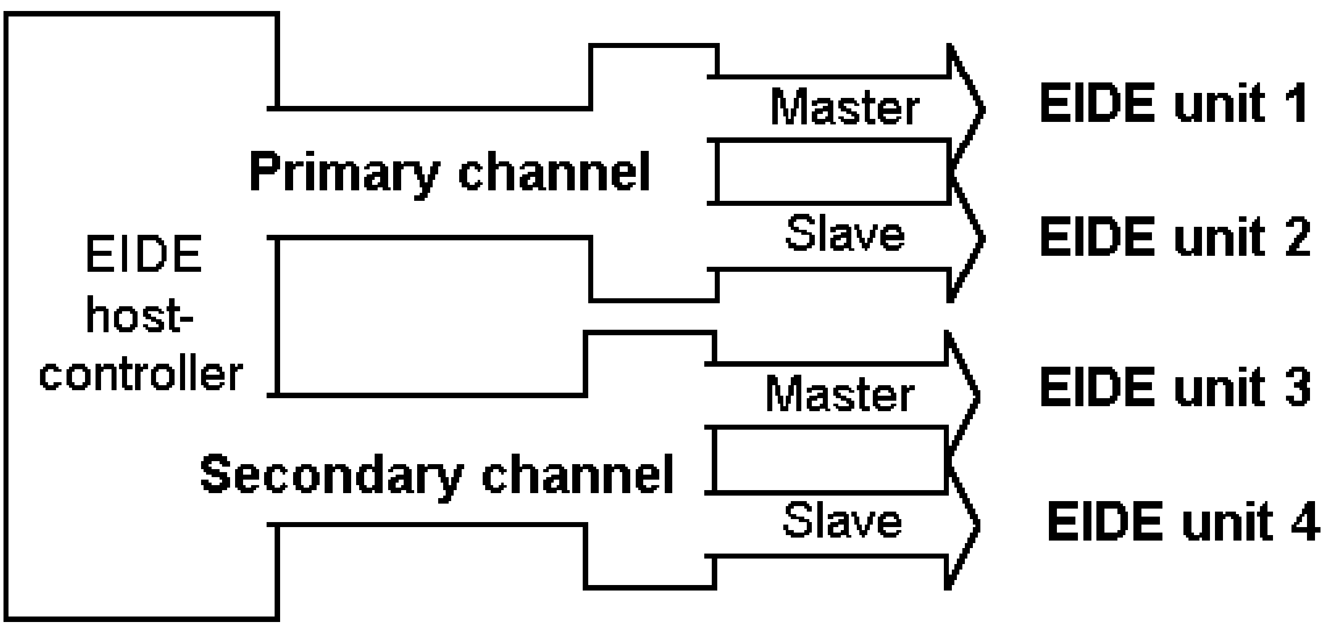
The connection

Here you see an EIDE cable:



Since each channel can handle two units, there are two of these connectors on the system board. Note the blind hole in top center. Note also the stripe (it is red) in the far right edge of the cable. It tells you that lead number one is on this edge. Both of these features help prevent incorrect installation of the cable.

The system board has sockets for two EIDE cables. Each EIDE cable (primary and secondary) has sockets for two units (master and slave).



If we use all four connections, it could look like this:

EIDE connection	**Unit**
Primary, master	Hard disk 1
Primary, slave	Hard disk 2
Secondary, master	CD ROM drive or DVD drive
Secondary, slave	LS120 diskette drive

Typically, a PC has two EIDE units connected: the hard disk and the CD ROM drive. However, as you can see, other units can be connected as well.

Transfer speeds and protocols

EIDE exists with different protocols, like PIO 3, PIO 4, and UDMA. They are backwards compatible, therefore choose the latest. The different PIO *modes* are significant for the transfer speed.

EIDE units communicate according to a specific *protocol*. Here you see the three most common, of which Ultra DMA is the newest:

Protocol for EIDE interface	**Maximum theoretical transfer**
PIO 3	13.3 MB/second
PIO 4	16.6 MB/second
Ultra DMA (*ATA-33*)	33.0 MB/second

The two main channels (primary and secondary EIDE) can each always run on their own protocol. However, sometimes the slave/master channels can not. Be aware of this potential problem. Is there only room for one common protocol? In that case the "winner" will invariably be the slowest of the ones connected.

Therefore, it is important that you connect your hard disk to an EIDE channel, which only runs Ultra DMA.

Ultra DMA requires the installation of drivers. Windows 95 does not recognize Ultra DMA. ASUS provides an excellent, simple patch program on CD. You run it just once. Then the drivers are stored in the right locations. After one or two re-boots everything works.

Another problem can arise if you connect two hard disks to the system board. Despite the suppliers assurance that "it is very simple," it does not always work. Therefore, it is important to start with one sufficiently large hard disk.

What does Ultra DMA offer?

The latest EIDE improvement is called Ultra DMA or Ultra ATA. It is a new interface, patented by Quantum, but supported by all system board and disk drive manufacturers.

The technology involves an improvement of the interface - the governing electronics which deliver the hard disk data to the system board. Quantum succeeded in eliminating the bottle neck in existing electronics to deliver data to the EIDE hard disks. The hard disk is no faster, but the data paths have been optimized.

On the surface, it sounds exaggerated to talk about a 33 MB per second transfer rate, well knowing that no EIDE hard disk can actually deliver more than 7 MB per second. Then PIO4, which can move 16 MB per second, should suffice? No, not so - the secret is in the the EIDE host controller. That controller, among other things, must retrieve data from the drive and deliver them to the PCI bus. Or it must retrieve data from the bus and deliver them to to the disk.

The host controller has certain administrative duties to handle between reading to/from the disk. And they take some time. One clock cycle in the EIDE controller lasts 400 micro seconds. Of these, 275 are spent on "administrative overhead" - handling commands, etc. The remaining 125 micro seconds are used to read from the hard disk. Therefore, a maximum transfer rate of 33 MB per second is necessary to keep up with the hard disk's capacity!

Actual measurements show that Ultra DMA disks yield up to 40% better performance than comparable PIO 4 disks. That is a clear improvement - even though the disks can never deliver the advertised 33 MB per second.

The system board and with that the chip set, must be set up for Ultra DMA in order for you to utilize such a disk. As always, check the chip set, when you buy a new PC. Since it provides solid performance improvement at no extra cost, it is important that it supports Ultra DMA.

Conclusion

For the EIDE hard disk to function in the Ultra DMA protocol, the following conditions must be met:

- The hard disk must be the Ultra DMA type.
- The system board must have a chip set, which supports Ultra DMA, such as 82430TX or 82440LX.
- BIOS must "log on" the hard disk with Ultra DMA protocol. You can verify that in the start up screen.

- Drivers for the chip set must be installed in Windows 95.

Ultra DMA ATA/66

Intel and Quantum has announced the new Ultra DMA standard called ATA/66 giving a theoretical speed of 66 MB/sec.

The new system requires a new cable with 80 conductors. The 40 new conductors are used for grounding. This way the noise is reduced, and the bandwidth goes up. The new cables use the old 40-pin plugs. If you use a ATA/66 system with a 40-pin cable, the speed will automatic decrease to 33 MB/sec.

What is AGP?

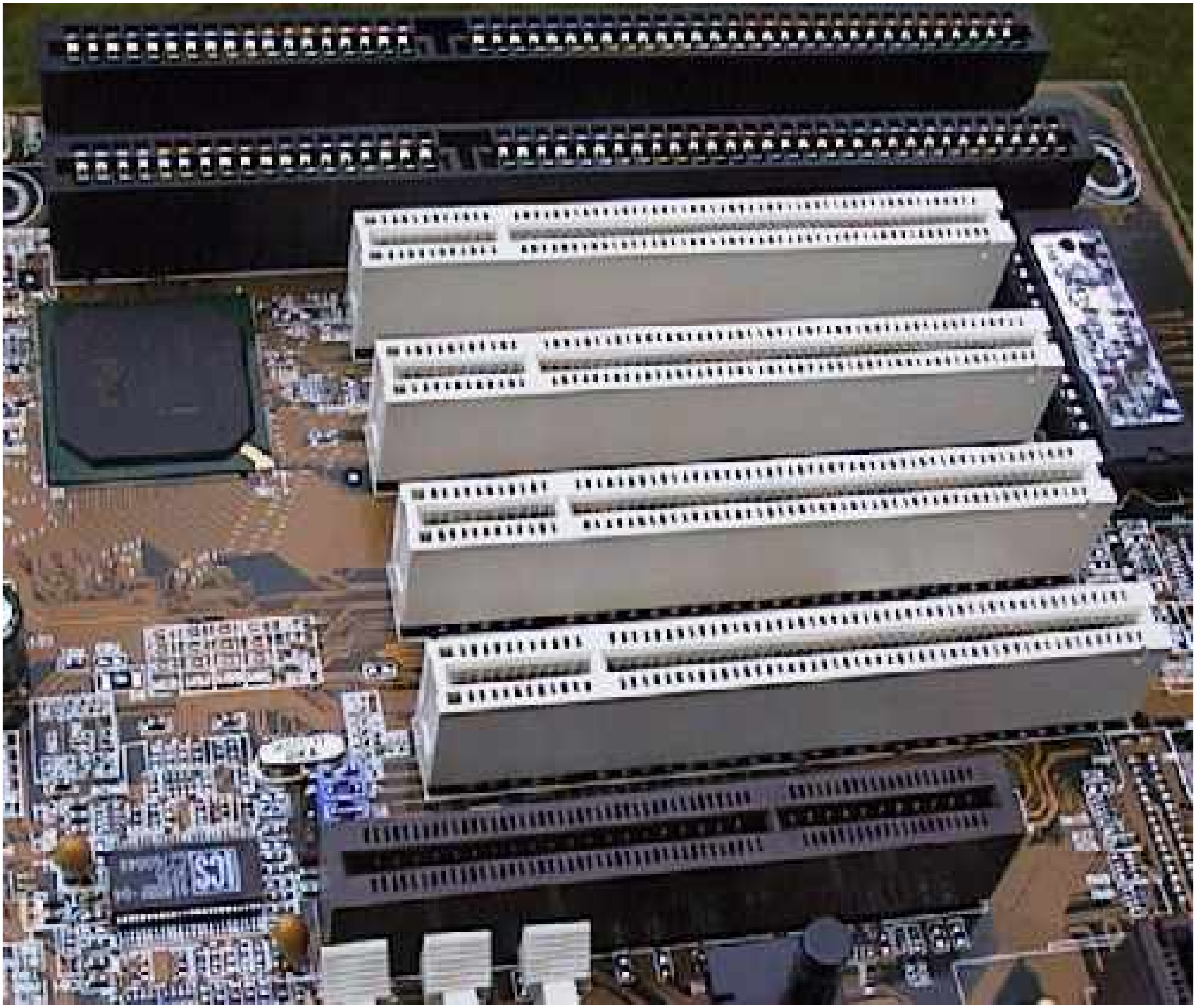
A new bus has arrived on the PC. It is called AGP (*Advanced Graphics Port)*. It is exclusively designed for video cards

AGP will probably not be in widespread use before 1998. Amongst other things, the system must be supported by the operating system (Windows 98). Likewise, it is claimed that the system bus will be raised from the current 66 MHZ to 100 MHZ, to allow AGP to prove its worth. AGP includes several techniques, of which two are understandable:

- PCI version 2.1 with 66 MHZ bus frequency. That is a doubling of transfer speed
- Possibility to utilize system board RAM for *texture cache*. This will reduce RAM card demand in connection with the most demanding programs.

One big AGP advantage is that the PCI bus is relieved of work with graphics data. It can concentrate on other demanding transport duties, like transfer to and from network adapter and disk drives.

Here you see the AGP-socket at the bottom. It looks like a PCI-socket, but it has been placed in a different position on the board. In the top you see two (black) ISA-sockets. Then four (white) PCI-sockets, and then the brown AGP-socket:



[About AGP]

About SCSI, USB and other serial busses

SCSI (Small Computer System Interface) is high end technology. It is a technology, which provide means for data exchange among hardware devices such as drives, tape streamers and scanners. SCSI is especially used in high end PC's such as network servers or just powerful workstations.

SCSI might be compared to the EIDE interface, which also uses a host adapter controlling drives. However SCSI have two major advantages to EIDE:

- A SCSI host controls 7 or 15 devices (using only one IRQ).
- The SCSI system holds its own computer power, thus freeing the CPU from workload.

If you are critical about your PC power, the SCSI would be worth considering.

The host adapter

A SCSI system is built around a central, intelligent controller called the *host adapter*. A host adapter can control several SCSI units:

- Many units on the same host adapter.
- Many types of *drives*: Hard disks, CD ROM's, MO drives like Zip drives, CD ROM recorders etc.
- Tape-streamers (DAT and others).
- Scanners.

The host adapter has its own BIOS separate from the PC's. When you boot the PC, you will see the adapter communicating with connected SCSI devices.

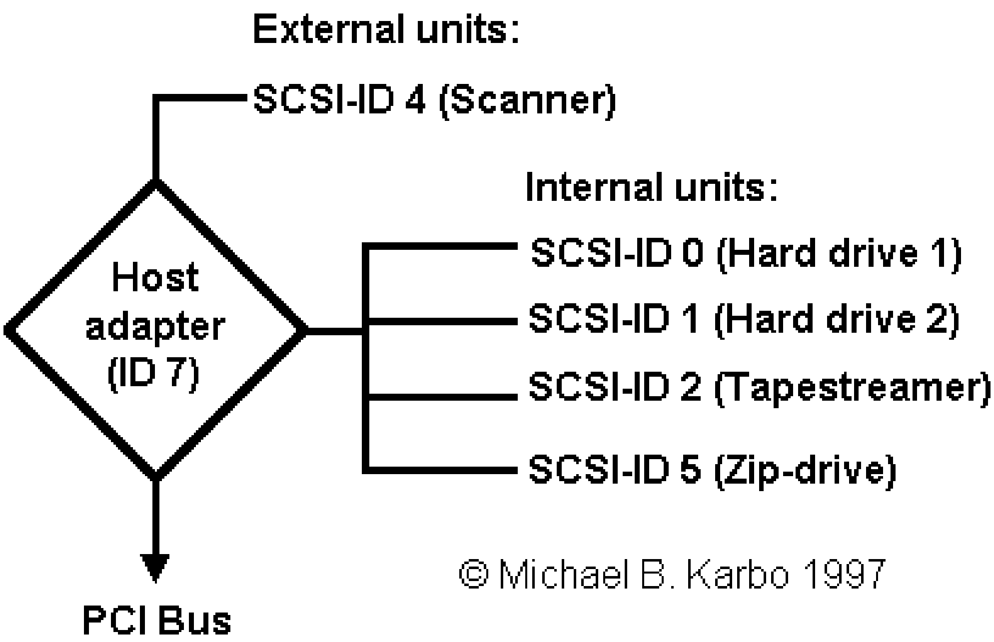
The adapter is rather expensive. Currently, the best for ordinary use is called Adaptec 2940 UW (priced at around $200). It is PCI based, so you could use it in your next PC too.

---

8 units in a chain

The regular SCSI 2 system can handle 8 devices (SCSI Wide handles 15). Each device has to be assigned an unique going from *ID 0* to *ID 7.* The SCSI devices can be *internal* (installed inside the PC cabinet) or *external*. The host adapter is a device itself. Typically, the host adapter will occupy ID 7.

Here is an illustration of a SCSI string with host adapter (ID 7) and five units (ID numbers 0, 1, 2, 4, and 5):

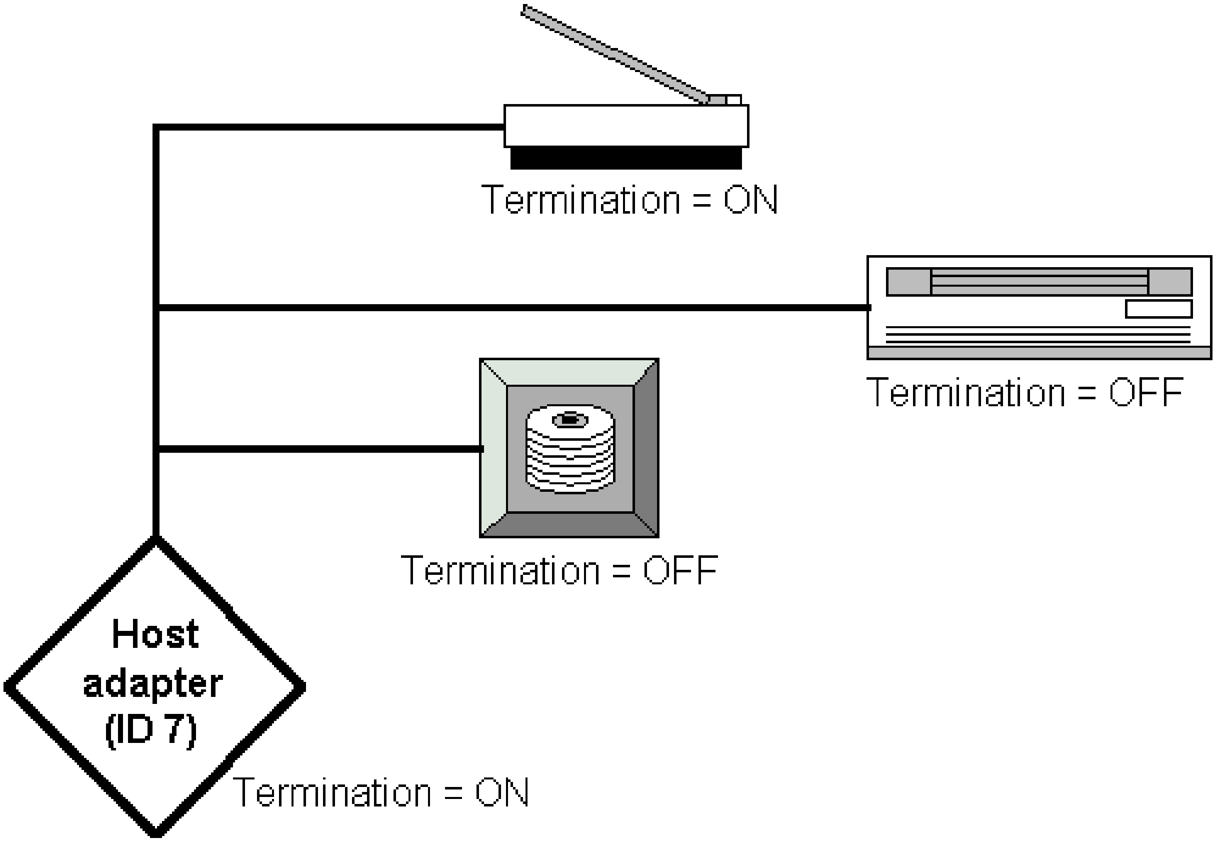


The total cable length in the SCSI chain must not exceed 6 feet.

Terminators in both ends

The last unit in both ends of the SCSI chain must be terminated. This means that there must be resistors (jumpers or switches) attached to two of the units.

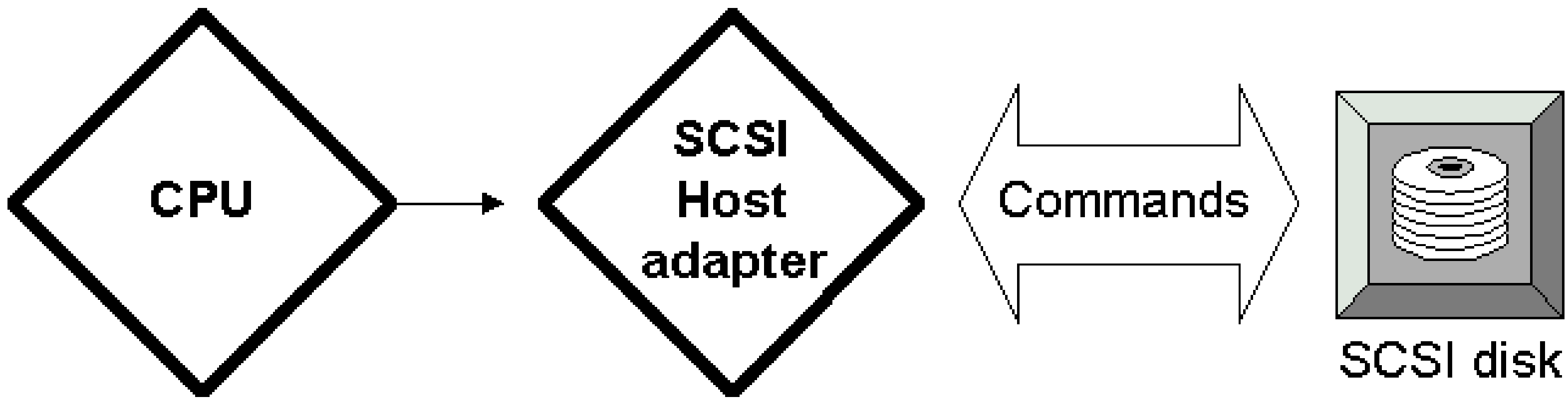
If you only use two devices, you do not have to worry about it. The host adapter is one end of the chain and the other device is the other end. With three or more units you have to take care of termination:



SCSI is intelligent

SCSI is remarkable in having an intelligent protocol, which assures maximum utilization of the data transfer. The basis of SCSI is a set of *commands*. Each individual device holds its own controller, which interprets these commands.

All commands within the SCSI system are handled internally, meaning the CPU does not have to control the process:



While the read/write head moves across a SCSI disk, the host adapter as well as the CPU can handle other jobs. Therefore SCSI is well suited for multitasking environments.

About the SCSI standard

SCSI stands for Small Computer System Interface. It is intended as a universal interface, defined and designed in 1982 by NCR and Shugart Associates. It exists in numerous variations. Here you see some of the more significant editions:

Standard	**Year**	**Bus speed**	**Bus width**	**Max. bandwidth**
SCSI-1	**1986**	**5 MHZ (Asynchronous)**	**8 bit**	**5 MB/sec**
Fast SCSI-2	**1990**	**10 MHz (Synchronous)**	**8 bit**	**10 MB/sec**
Fast-Wide SCSI-2	**1992**	**10 MHZ (Synchronous)**	**16 bit**	**20 MB/sec**
Fast-Wide Ultra SCSI-20	**1994**	**20 MHZ (Synchronous)**	**16 bit**	**40 MB/sec**
SCSI-3	**1996**	?	?	**80 MB/sec**

SCSI-2 is the 16 bit standard from 1990. It is found in Fast and Fast-Wide-SCSI. Today. there are many SCSI standards. Among others, you can come across SCSI-20 and SCSI-40, which refers to the bus speed. The last one is also called SCSI-3, but its standard is not finalized.

The SCSI standard seem to have its own life with plenty of new development.

What do you gain with SCSI?

Expensive but good: SCSI makes the PC a little more expensive, but much better. That's all. The advantages are, that on the same PC you have free access to use *many* units and *good* hard disks:

- **It is easy to add many high end accessories, such as DAT-streamers, CD-ROM recorders, MO drives, scanners, etc.**
- **You can use SCSI hard disks.**
- **You can use CD-ROM drives on SCSI , where they perform a lot better than on IDE.**

Advantages of SCSI hard disks

SCSI hard disk are generally of higher quality than other disks. EIDE disks come in various qualities from different manufacturers. However, even the best EIDE disks cannot compete with the best SCSI disks.

Typically, good SCSI disks come with a 5 year warranty. They come in larger capacities than the EIDE disks and they are faster. At 5400, 7200 or 10.000 RPM they have shorter seek times. They also have a bigger cache.

Another advantage is the large number of accessories, which can be attached. If you buy a 4 GB SCSI disk today, you will guaranteed need additional disk storage in a few years. Then you just add disk number two to the SCSI chain, and later number three. The system is more flexible than EIDE, where you can have a maximum of four units incl. CD-ROM.

The SCSI hard disks can also adjust the sequence in the PC's disk read commands. This allows to read the tracks in an optimal sequence, enabling minimal movements of the read/write head. Quantum calls this technology ORCA (*Optimized Reordering Command Algorithm*). It should improve performance by 20%.

Finally, the SCSI controller can multitask, so the CPU is not locked up during hard disk operations, which you can experience with IDE.

SCSI hard disks can achieve substantially larger transfer capacity than the IDE drives, but they have the same bottle necks: the serial handling of bits in the read/write head, where the capacity is highly dependent on the rotation speed.

Booting from SCSI disk

If the hard disk has to be booted, traditionally it has to be assigned ID 0. If the SCSI controller has to control the hard disk, then the PC CMOS setup must be modified, so the (IDE) hard disk is not installed if not both types of hard disks are installed.

The operating system will find the host adapter after start up and BIOS will be read from the hard disk through the adapter. New BIOS's allow a choice of booting from either IDE or SCSI disk.

Fast and Ultra Wide:

The newest generation of SCSI hard disks are both fast, ultra and wide. Therefore, the best advice is to buy an adapter like Adaptec 2940UW2, which can handle the newest disks.

IBM disks:

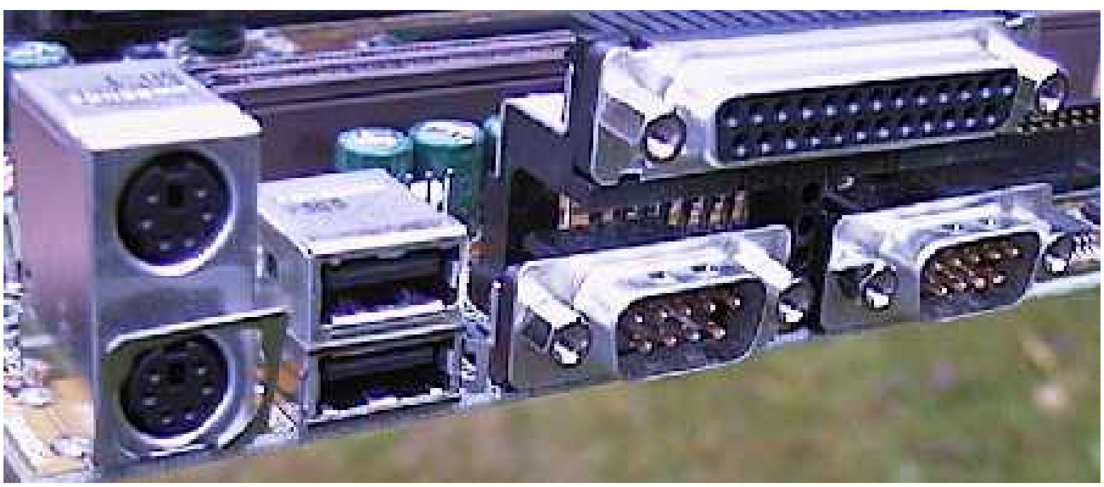
Allow me to advertise IBM's SCSI disks. They are fantastically good. Unfortunately, not many people know about them. I have had a few of them. They excel in high quality at reasonable prices. The physical construction is very appealing: The electronics are integrated in very few components. Everything exudes quality! And they are very quiet. You simply can not hear them.

32 bit problems in Windows 3.11

Windows 32 Bit Disk Access has given problems with SCSI disks. For a long while, it was impossible to install a 32 bit driver in Windows 3.11 to the SCSI disk. This was solved in 1995 and there are no problems with Windows 95 and NT.

USB

The USB (*Universal* or *Useless Serial Bus*) is a cheap, slow bus running up to 12 Mbit/sec. Just as FireWire it is an open royalty-free specification. The USB holds up to 127 units in one long chain. Units can be plugged an unplugged on the fly very easily. Here you see the plugs, the two small ones, number two from the left:



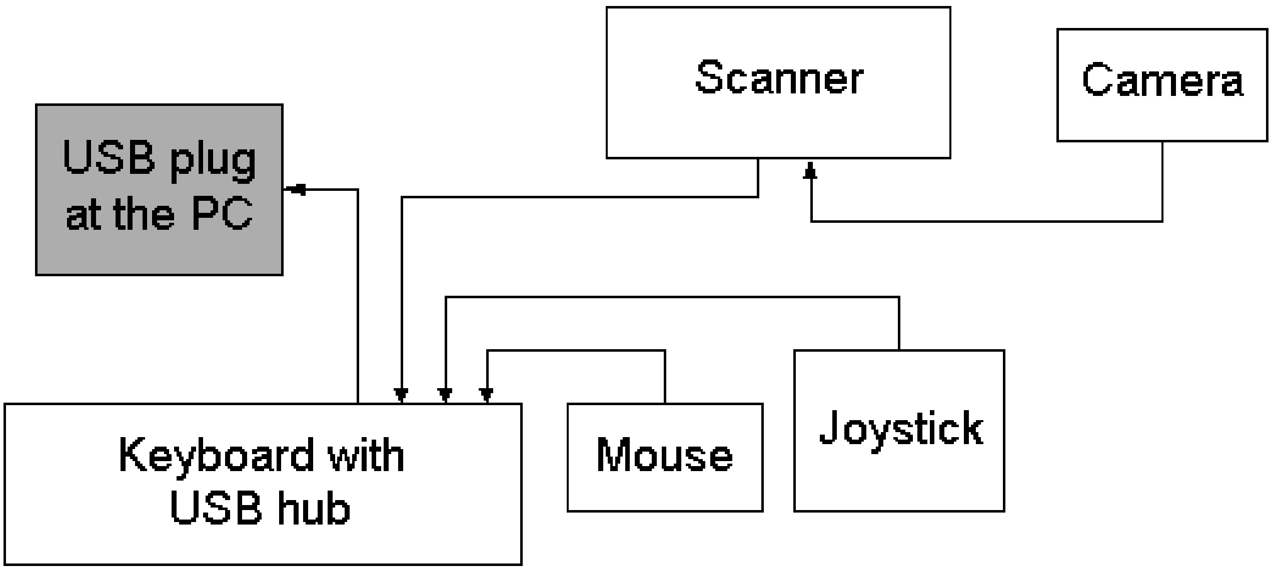
There will be problems with USB in the beginning, since many motherboard manufactories produced their own versions of the port before at was fully standardized.

USB is only supported by Windows 95 OSR2.1, but with Windows 98 we shall really use it. USB will replace the mess of cables and plugs we today use for:

- Keyboard
- Mouse
- Joystick

- Loudspeakers
- Printers
- Modems and ISDN-adapters
- Scanner and camera

All these units - and lots of others - will be connected using one single plug at the PC. The keyboard may hold a hub, so other USB-units is connected here. Each unit holds at minimum two plugs, so they all can be daisy chained:



All units have a firmware identification code, that communicates with Windows 98. The unit must have a power feed (could be minimum 100 ma) to be recognized by the USB controller and Windows 98. If one unit fails this way, Windows shows an ! on yellow background to signalize that something has to be done. This could be to unplug other USB devices to increase the available power in the chain.

Many hardware manufactories today produces their modems, cameras and scanners in versions with two way interface. The device connects traditionally using a COM-port - or you use the USB.

The Hub

We should be able to connect 127 units all together. An important Unit is the hub, as we know it from the ethernet. The USB hub may be found in the keyboard, but probably we will use little, powered boxes holding 8 USB connectors. Five of these hubs can be daisy-chained, providing connection for 36 units. USB hubs can for convenience be placed on the backside of monitors, in scanners aso.

The USB cabling can deliver 500 mA of power. This sufficient to feed a keyboard or other low-powered units. But it is not enough for multiple units, therefor we will need powered hubs.

We shall also find COM to USB converting hubs. A box will house four DB9 connectors serving as COM5, 6, 7 and 8. They all connect to the PC via one USB port. This way serial devices can connected without the IRQ puzzle we often experience nowadays.

Shared USB-units

An other interesting aspect is that USB allows shared peripherals. This means that two PC's can share an USB-unit. Or you may even use the USB for a low-priced network connection?

IEEE 1394 FireWire

The next technology is called FireWire. It does not look very much like the SCSI we know, but is a further development being a serial high speed bus.

The interface IEEE 1394 has a bandwidth of 400-1000 Mbits per second. It handles up to 63 units on the same bus. The units can be plugged and unplugged *hot* - meaning you do not have to power-down the PC.

The Firewire is expected to replace:

- Parallel Centronics port
- IDE
- SCSI
- EIDE (later on)

Entertainment

The first versions will be used for digital audio/video-electronics like:

- Digital cameras and cam-corders ao.
- DVD drives
- Scanners

FireWire comes from Apple but it is an open standard which can be used for free. Hence all mayor hardware companies has adapted it. Especially the entertainment electronic industry (Video/games/television) have great hopes with FireWire. It will connect all types of digital electronics with the PC and this way open up for a much more modular design.

Since FireWire is advanced and yet claimed to be cheap & simple, the communications protocol can handle a lot of other units like:

- Network controllers
- Hard disks, CD-ROM drives
- Printers

Two modes

The FireWire standard operates with two *modes.*

- *Asynchronous* as other busses. This means that operatings across the bus is controlled using interrupt signals. The bus reports to the host when a task is fulfilled.
- *ISO synchronous*. In this mode data is being transferred at a steady preset speed - continuously and without any supervision from the host. This opens up for data-streaming useful for video or the multimedia presentation.

The FireWire is an peer-to-peer interface. This means that data can be transferred between two units attached to the bus without supervision from the PC.

FireWire has 64 bit address bus. Compared to SCSI each unit does *not* need an unique ID, they are dynamically configured "on the fly". Neither does the bus have to be *terminated*. All together a lot more simple than SCSI.

One of the problems with SCSI has been the limitation on distance between the units. FireWire can hold up to 16 units in the same "string" and there can be up to 4,5 meters between two units.

The first implementations of FireWire will connect it to the PCI bus using the new PIX6-controller, which will be a part of one of Intel's new chip sets. I think it will last 2-3 years before we really see this new technology in the market. But it will be worth waiting for it, it opens up for new world of inter connectivity between TV, PC, video end all other types of electronically gear.

USB and FireWire - serial busses of the future

In the future a PC will hold only two I/O busses - both serial:

- USB for all low speed gear.
- FireWire for high speed I/O as disks, video ao.

With FireWire and USB motherboard and software configuration will be vastly simplified. I imagine one driver for each bus covering all units on the bus. No setup of IRQ, DMA and I/O-address - great. The PC will thoroughly become a modular setup of Plug And Play units!

Links You find technical specifications etc. in these sites:

About SCSI.

Apple about: FireWire

The IEEE has its own page at http://www.ieee.org.

Intel's USB site http://www.intel.com/design/usb

USB site: http://www.usb.org

Click & Learn. Module 6a. WWW.MKDATA.DK

About file systems: DOS formatting, FAT, etc.

We have seen before that the PC is a big data processor. We have also seen that data are bits and bytes, which are organized in files

One of the operating system's major tasks is to write these data to a disk. Hard, floppy, and zip disks must be formatted before we can save files on them. In these pages, we will review formatting, file systems, etc. We will start with a general view, then go in depth about FAT formatting, which is (still) the most common.

What are file systems?

As I wrote in module 4a, drives are storage media, which can hold a file system. When a disk is formatted in a drive, it becomes organized and prepared to receive data. When we format a disk, it receives a *file system*.

Formatting can be compared to starting a library. You must install the book shelves and the catalogue system before any books are put in place. Once the library is ready, bring on the books! Similarly with a disk. When we format it, we "burn in" a file system to make it ready to receive data (files).

We can format with any one of several different file systems:

- **FAT (*File Allocation Table*),** the original, old 16 bit DOS system is probably used in 90% of all PC's. It is also called FAT16 contrary to:
- **FAT32 a new addition to FAT**, which Microsoft introduced with Windows 95 B – The December -96 version (OSR2).
- **HPFS (*High Performance File System*)** from OS/2. It is an advanced 32 bit file system, which in all respects is far superior to FAT, except for possible usage. It can only be used with OS/2.
- **NTFS from Windows NT**. A 32 bit file system like HPFS, but *not* compatible with it. NTFS can, unfortunately, only be used in Windows NT. If it was available for use in Windows 95, it would be far to preferable to FAT and FAT32.

- **NetWare is a server operating system from Novell.** It has its own 32 bit file system. For that reason, the Novell server, contrary to NT or OS/2 servers, cannot be used as a work station. The file system is much faster than FAT, but it works only with Novell servers (typically file servers).
- **UNIX** servers have their own filing system. Here the use of upper/lower case in file naming is significant.

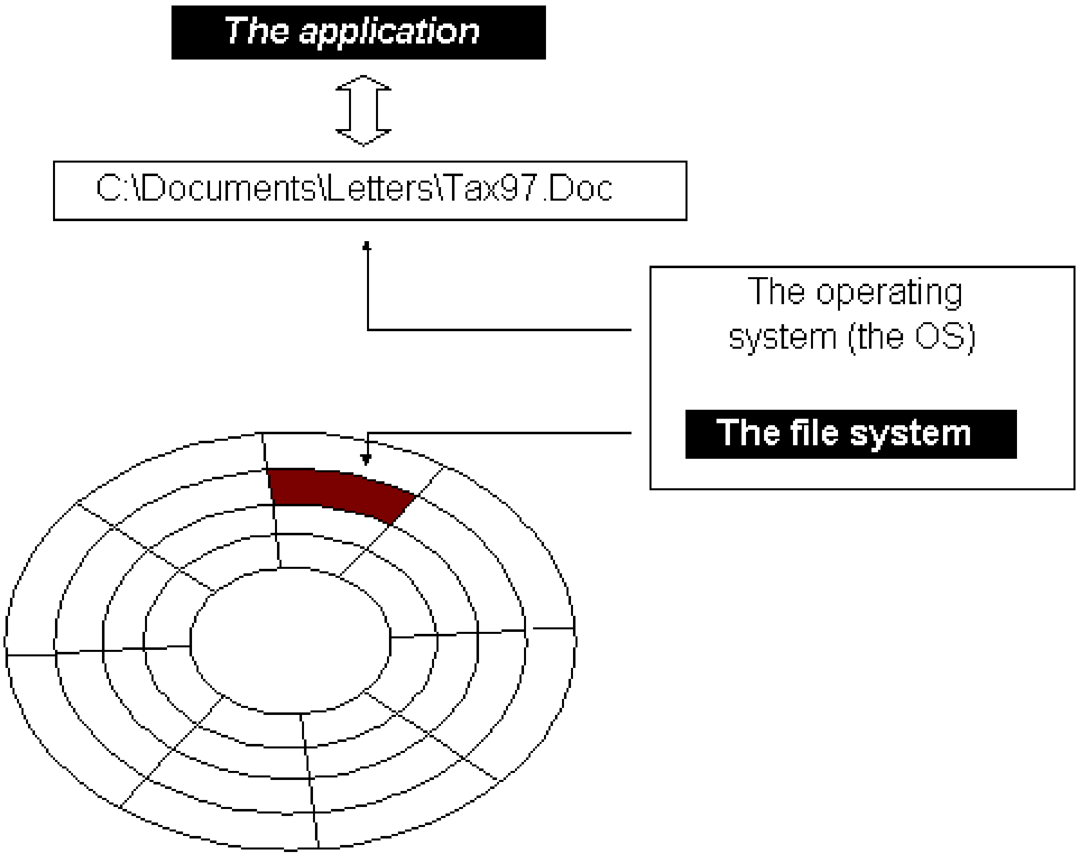
Read in the following pages about the concepts of these file systems.

Relationship between file system and operating system

We see that that the file system is an integral part of the operating system. An operating system can sometimes work with different file systems:

Operating system	File system(s)
DOS	FAT16
Windows 95	FAT16, FAT32 (actually VFAT: *Virtual FAT,* which provides access to use long file names)
Windows NT	FAT16, NTFS
OS/2	FAT16, HPFS
Novell NetWare	Own "brand name"

The file system is actually the *interface* between operating system and drives plus disks. When the user software, such as MS Word, asks to read a file from the hard disk, the operating system (Windows 95 or NT) asks the file system (FAT or NTFS) to open the file:



The file system knows where files are saved. It finds and reads the relevant sectors and delivers the data to the operating system.

Limitations in disk size

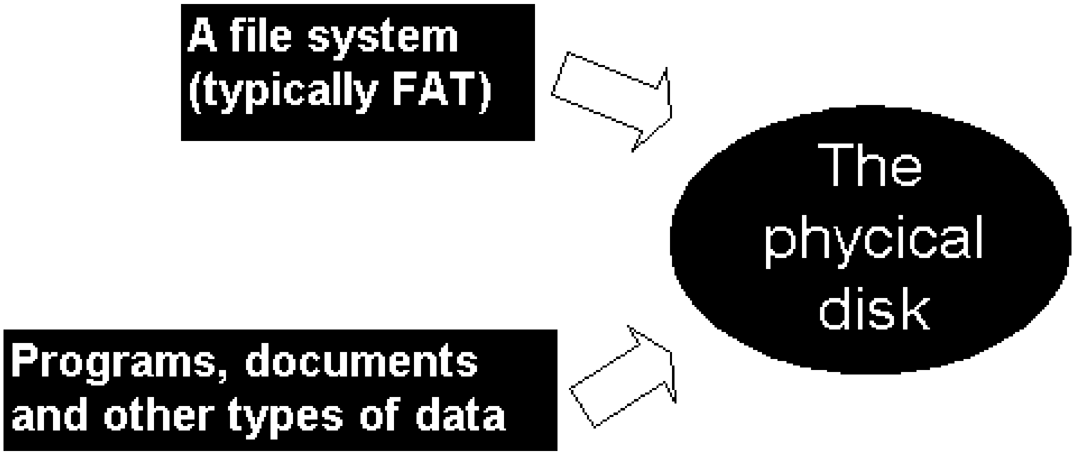
Over the years, the PC has suffered from a long list of irritating limitations. The hard disk industry has continuously developed hard disks with increasing capacity. However, the system software (BIOS, DOS, and FAT) has set its limitations:

- DOS versions below 3.0 could only handle hard disks up to 16 MB.
- Versions 3.0 to 3.32 could handle up to 32 MB.
- DOS 4.0 could handle up to 128 MB.
- DOS version 5.0 and the BIOS, which controls IDE drives, could only accept 1024 cylinders and disks up to 528 MB. This limit was broken with the EIDE standard.
- Today, FAT16 can handle a maximum of 2 GB because of 16 bit calculations of the cluster size.
- FAT32 accepts disks up to 2048 GB. This standard will probably last another a couple of years.

Let us return to the file system:

The formatted disk

We know that a disk must be formatted with a *file system*, before it can accept files to be saved:

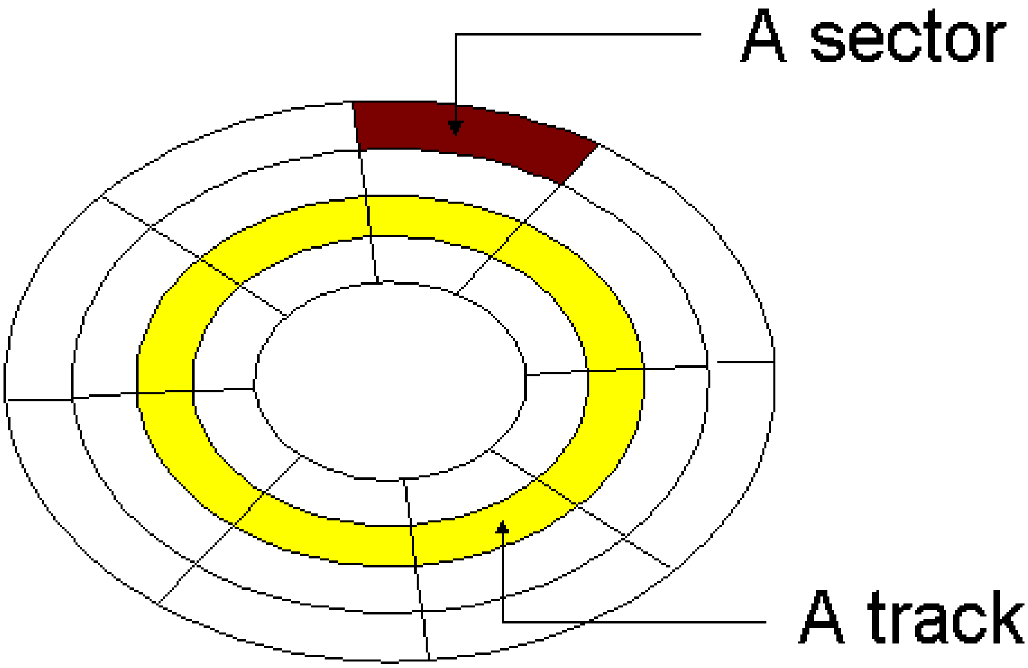


Now let us examine the disk formatting process. How does it work?

About sectors

All disks are divided in 512 byte sectors. That is the standard size for the smallest disk unit. You could easily format with a different sector size, but that is not done. A sector is then the smallest disk unit and it holds 512 bytes of data.

Sectors are created when the circular disk is organized in concentric tracks. Each track is divided into sectors. Each sector can hold 512 bytes.



But, how are these sectors distributed? How are the files placed in the sectors? How do we handle a file larger than 512 bytes, which must occupy more than one sector? Who keeps track of what is in each sector?

This is a task for the file system. Below, we evaluate hard disks only and only FAT. Despite its age and flaws, it is still by far the most widely used file system. As for diskettes, read about diskette formatting.

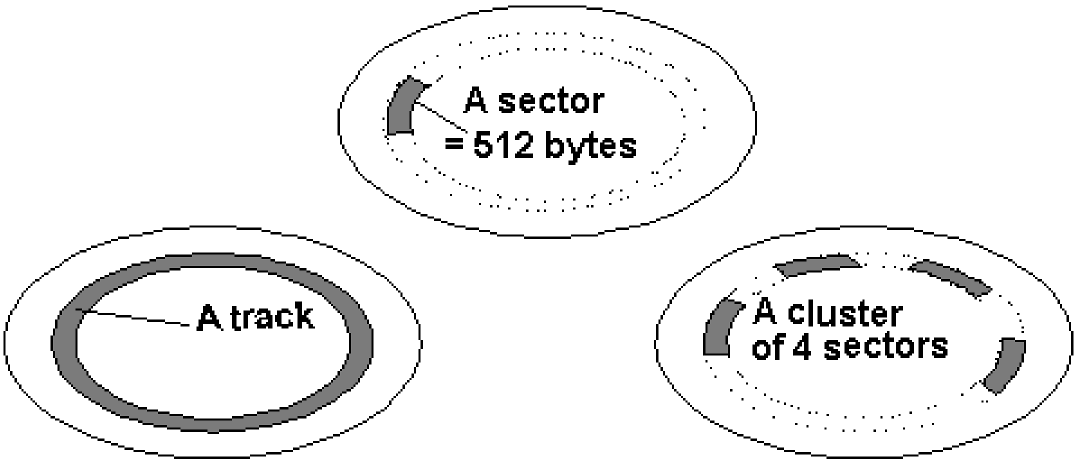
About clusters

To continue in the understanding of the file system, we must introduce a new concept - *clusters*.

Each sector holds 512 bytes and a sector is the smallest disk unit. However, often a sector is too small for DOS to handle. That is because DOS is a 16 bit operating system.

By design, DOS can only handle 2^{16} disk units at a time. A disk unit (my expression) is either a sector, or a cluster of sectors. Thus, DOS can only handle 65.536 of those!

Therefore, in FAT formatting the sectors are gathered in *clusters* of 2, 4, 8, 16, 32, or 64 sectors:



The cluster concept is an administrative invention. They are necessary, to allow DOS to handle large disks.

They are also called *allocation units*. The number of sectors gathered in one cluster depends on the disk size:

Disk size (partition size)	**Cluster size**
< 255 MB	**8 sectors (4 KB)**
< 512 MB	**16 sectors (8 KB)**
<1024 MB	**32 sectors (16 KB)**
< 2048 MB	**64 sectors (32 KB)**

In Dos, the data area of the hard disk is divided into a specified number of clusters, which of necessity increase in size with the size of the disk. On modern hard disks, the clusters will usually be 16 or 32 KB, as illustrated above

---

Small clusters with FAT32

The good news is that FAT32, found in the latest version of Windows 95 (B), handles disk formatting much better than FAT16. With FAT32 it is possible to format hard disk *partitions* of more than 2 GB with small cluster sizes:

Partition	**Cluster size**
<8 GB	4 KB
8 GB - 16 GB	8 KB
16 GB - 32 GB	16 KB

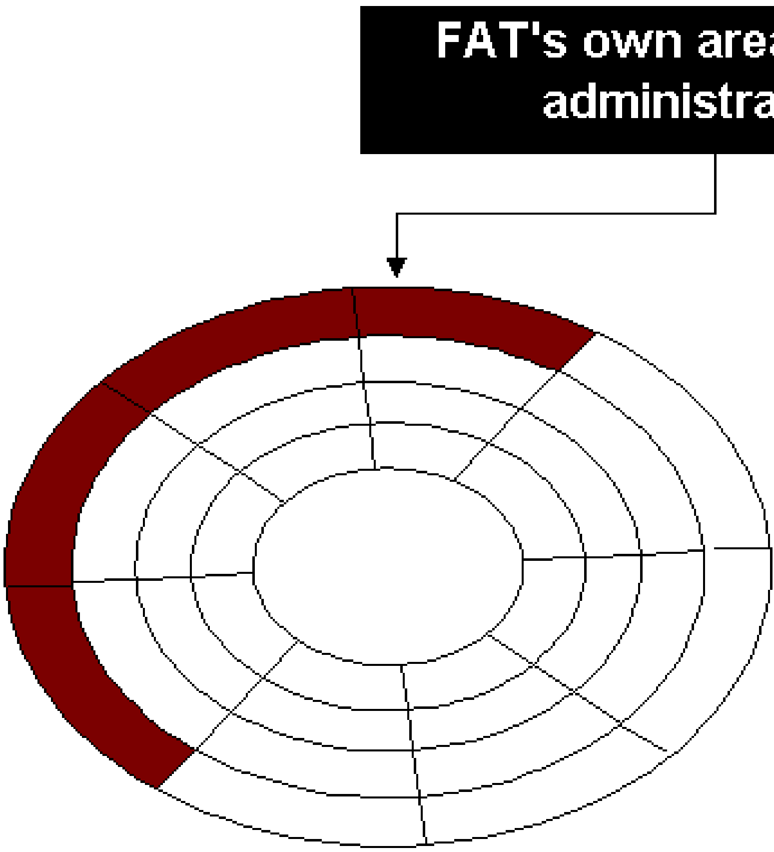
>32 GB	32 KB

Something else new in FAT32 is the movable root directory, which can be of variable size. It involves active use of both FAT's (I cannot explain how). Altogether, it should make it simpler and safer to change partition sizes. But the number of clusters per partition grows enormously in large partitions.

FAT32 can only be installed in a new PC, since the partition has to be formatted in a special manner. The file system is only available in the Windows 95 B version, also called OSR2 (OEM Service Release 2).

The FAT formatted disk

During formatting, all hard disk are divided into multiple sectors. The sectors must contain *both* user data and the file system administrative data. This is because in FAT, the administrative data are stored on the disk also:



Thus, the disk is divided in:

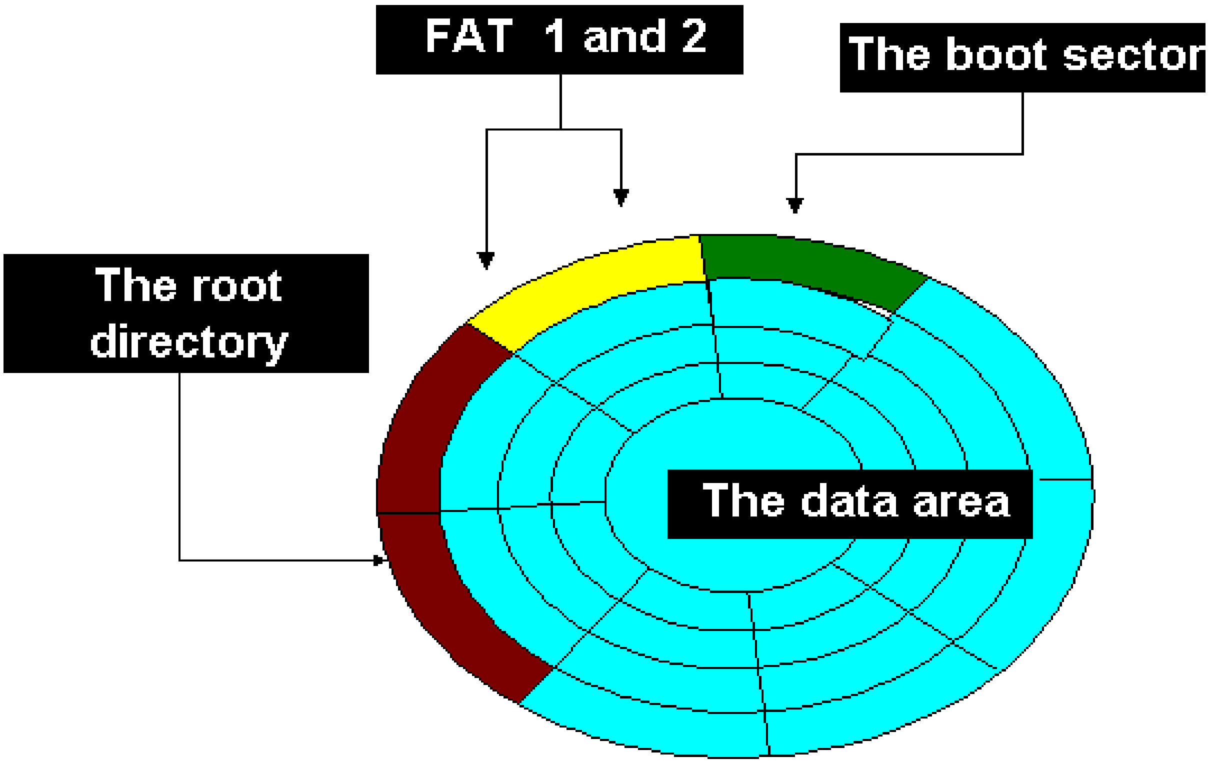
- **Sectors, occupied by FAT administrative data.**
- **Sectors, which are user available for data storage (the data area).**

The four disk areas

Each disk or disk partition contains four fundamental areas:

- **The boot record, which is always in the *first sector*.**
- **FAT areas, of which there are usually *two* identical.**
- **The root directory.**
- **The data area, where all files and sub directories, beyond the root directory, are stored. The data area sectors are gathered in *clusters.***

This organization is illustrated here:



An example - the sectors in the four areas

Let us look at a FAT formatted hard disk with 160 MB: How are the sectors utilized?

The partition contains exactly 157.9 MB, if we calculate correctly. We are talking about 165,656,576 bytes. The total data storage area is divided in 323,548 sectors, 512 bytes each. If you multiply, that results in 165,656,576 bytes.

323,548 sectors		
512 bytes each	**= 165,656,576 bytes**	**= 157,9 MB**

The file system now assumes control over these 32,548 sectors. The boot record occupies the *first sector*. Here is a brief description of that and other administrative areas:

Boot record

The first disk sector is always reserved for the *boot record.* It contains information about the disk or its partitions. A physical hard disk can be divided into different partitions. DOS, Windows 95 and NT treat each partition as a separate drive.

The boot record information enables the file system to handle the disk. At the same time, it includes a *small* program to use during system start-up. Here is a summary of that sector's contents (skip, if you do not understand):

8086-instruction (JUMP).
DOS name and version number.
Bytes per sector.
Sectors per cluster in the data area.
Number of reserved sectors.
Max. number of entries in the root directory.
Total number of sectors.
Media description (is this a hard disk?).
Number of sectors per FAT
Number of sectors per track.
Number of disk read heads.
Number of hidden sectors.
BOOT-strap program routine, which reads the hidden file (like IO.SYS), which starts the operating system.

The boot record is found on all disks, regardless of whether they are FAT or otherwise formatted. That sector contains the necessary description of the partition.

The FAT areas

After the boot record, we get to the FAT areas. There usually two identical FAT's. FAT number 2 is simply a spare copy of number 1, since FAT is essential for the function of the disk.

The FAT file administration is actually a very simple system, but it is complicated to describe. Later, I will show some practical examples. Here is the first description. Even if you do not entirely understand the following, do not give up.

FAT consists of a table of whole numbers, which has 65,536 16 byte *entries*. Each of these entries contain information about a cluster.

The content of each FAT entry consists of a *whole number*. In the table below, they are written as four digit hexadecimal numbers, which show one of four options.

Possible FAT cluster entry	Value
The cluster is part of a file, the last in the file.	**FFFF**
The cluster is part of a file. You can read the number of the next cluster in the same file.	**like A8F7**
The cluster is empty, thus free.	**0000**
The cluster contains defective sectors.	**FFF7**

Example on reading a file

When the file system has to read a file, it follows this routine. We imagine that the file occupies 4 clusters and it occupies cluster numbers 442, 443, 444, and 448. But how does the operating system read these addresses?

- Find the file directory entrance (from its file address)
- Read the first cluster, number 442 in the directory entrance
- Look up in FAT under number 442. We find the number of the next cluster (443)
- Look up in FAT under number 443. We find the number of the next cluster (444)
- Look up in FAT under number 444. We find the number of the next cluster (448)
- Look up in FAT under number 448. Here is the number FFFF. That was the last cluster.

FAT always works in this way. Whenever a file has to be read, its location is read in the table. Every time a file has to be written to a disk, vacant clusters must be found for it, and the information is stored in FAT, to facilitate retrieval.

One of the great advantages of disk cache programs are, they always have a copy of FAT in RAM. In this way the disk cluster "map" can be read much faster than if the operating system had to read the FAT from the disk at each request.

The size of FAT

Since each cluster has a FAT entry, the size of the FAT areas depends on the disk size. Each entry occupies 16 bytes. Let us return to the sector account in the example of a disk of 160 MB size:

The maximum FAT size is 128 KB, since 2^{16} files, 2 bytes each, equals 65,536 x 2 = 131,072 bytes or 128 KB. In our example, there turns out to be 40,400 *clusters*, since the disk partition is 160 MB.

We have two FAT's, at 40,400 X 2 bytes. That comes to a total of 161,600, and that will occupy 316 sectors.

The root directory and other directories

The last administrative area on the disk is the root directory. Since there are always 512 file or directory *entrances* in the root directory, it is the same size on all hard disks. The root directory is unique in its fixed size and its location in the root. Other than that, it is a directory like any other.

Actually, a directory is a list of files and other directories. Thus, you can read the names of files and sub directories in the directory! The directory structure consists of a number of directory entries.

Let us look at these directory entries, each of which occupies 32 bytes. The directory entries are identical, whether they are in the root directory or a sub directory.

These entries, 32 bytes each, contain a lot of information like:

- The file name (in 8.3 format)
- File size in bytes
- Date and time of last revision

You can see the layout of the file entry on the illustration to the right. The 32 bytes are grouped in sections. This holds true for *all* entries, whether they point towards files or directories. This holds true for the root directory as well as all sub directories.

Note that we also find the number of the first cluster. This is important, because this is where the operating system starts to localize the file.

Remember the description of FAT above. You see that the start cluster number is read in the directory entry for the file.

Next FAT reads the numbers of cluster number two and so on, if the file is spread over additional clusters.

The location of any file is described in this manner: The first cluster is read in the directory entry (root or sub directory). The following cluster numbers are retrieved from FAT.

On FAT16 formatted hard disks, the root directory occupies 512 entries, which are 32 bytes each. Thus, it occupies 16 KB.

File name 8 bytes
Extension 3 bytes
Attribute 1 byte
Reserved 10 bytes (FAT32 uses two of them)
Time 2 bytes
Date 2 bytes
First cluster 2 bytes
File size 2 bytes

All sub directories have at least two entries. They are rather *special*, in that they refer to the directory itself and to its "parent" directory (in which it is a sub directory). The entries can be seen with the DOS command DIR.

The entry for the directory itself is seen as one dot. The entry for the parent directory is seen as two dots.

```
.             <DIR>
..            <DIR>
```

The Data area

The rest of the *disk* contains the most important part, the data area, where all files and sub directories are stored. The data area is by far the largest part of the disk.

The sectors in the data area are *allocated* in clusters. As mentioned before, the maximum number of clusters for data is 2^{16} = 65,536. Our hard disk is 160 MB. That results in 40,400 clusters, 8 sectors each.

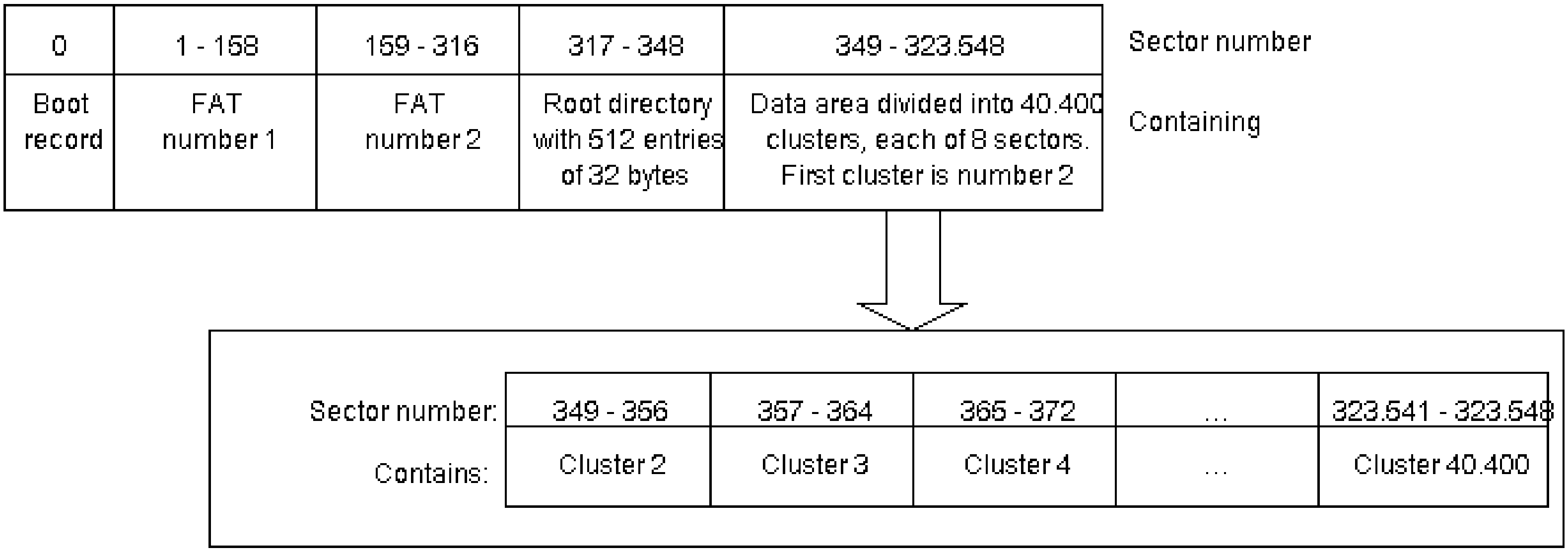
All sub directory entries in the data area are organized in 32 byte files, which contain the same fields as the root directory entries.

Completing the account

It took long explanations to bring us to here. Now let us pull it all together. The user has a 160 MB *hard disk* , but that is a somewhat theoretical view. Actually, the disk contains 323,548 sectors, 512 bytes each. They are distributed like this:

Area	Number of sectors	Sector number
Boot - record	**1**	**0**
FAT 1	**158**	**1 - 158**
FAT 2	**158**	**159 - 316**
Root directory	**32**	**317- 348**
Data area with 40,400 clusters of 4 KB	**323,200**	**349 - 323,548**

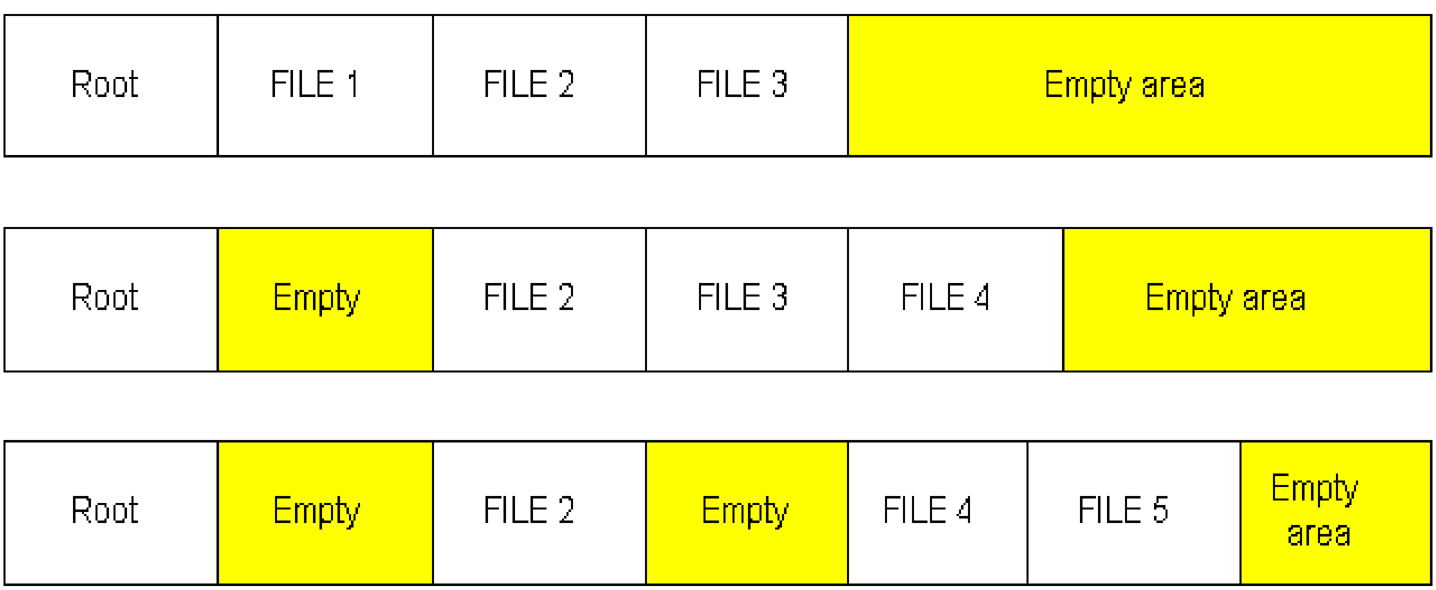
Here is a graphic illustration of the same distribution:



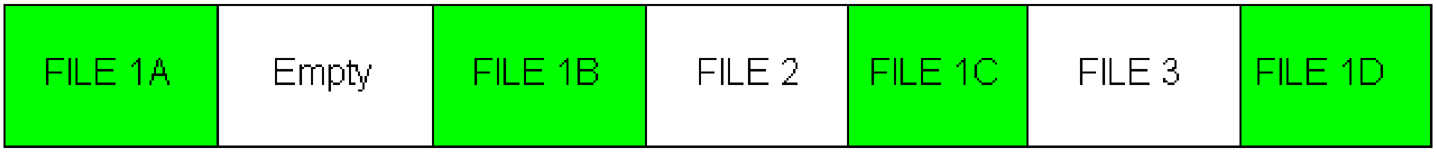
File fragmentation

When we work with FAT formatted disks, file *fragmentation* will occur all the time. One file can be several megabytes, thereby occupying more than one cluster. Maybe it requires 17 clusters. Ideally, the 17 clusters should be located next to each other. They can then be read at optimum speed, since that allows minimal movement of the read head. However, that is not the way it works.

In actual operation, the individual files are broken up in multiple blocks, which are scattered across the disk. The problem increases with time. The more files you have on the hard disk, the more fragmentation you will experience. To begin with, vacant spaces appear between the files:



When you first write to a new hard disk, the file might occupy 17 clusters in sequence. The same will happen to file number 2, 3, etc., until there are no more vacant clusters on the disk. Then the file system must re-use clusters. That is done by finding empty clusters, where the contents have been erased. Thus, the file could be scattered in 17 clusters, *none* of which are in sequence. Here you see a file split in four unconnected clusters:



In the first DOS versions, when a new file had to be written, the file system always returned to the first vacant cluster to start a new file. That was done, to get optimum utilization of the disk. It also resulted in immediate and total file fragmentation. Since DOS version 3.0 the system was changed to fill the disk, before any vacant clusters were re-used. That delays fragmentation, but sooner or later it will occur anyway.

Defragmentation

You can use the program DEFRAG to defragment the files on the disk. If you are a heavy PC user, it needs to be done often. I usually run **SCANDISK** first. That checks the file system for logical errors and repairs them. Scandisk will often find errors, so it does a good job.

Next defragment the disks with **defrag /all**. Both programs can be started with the command Start --> Run. Type in the command on the window: defrag /all

Here you see the defragmentation:

Run defrag weekly - that will keep your hard disks in good shape. Don't make the intervals too long. That can cause the disk to get messed up, especially if it is nearly full.

Partitioning with FDISK

Hard disks can be divided in more than one partition. That is done with the program FDISK, which is found in all PC's - regardless of which version of DOS, Windows, or OS/2 is the operating system. They all have FDISK.

FDISK can divide the hard disk in up to four partitions. In FAT16, the individual partition must not exceed 2 GB.

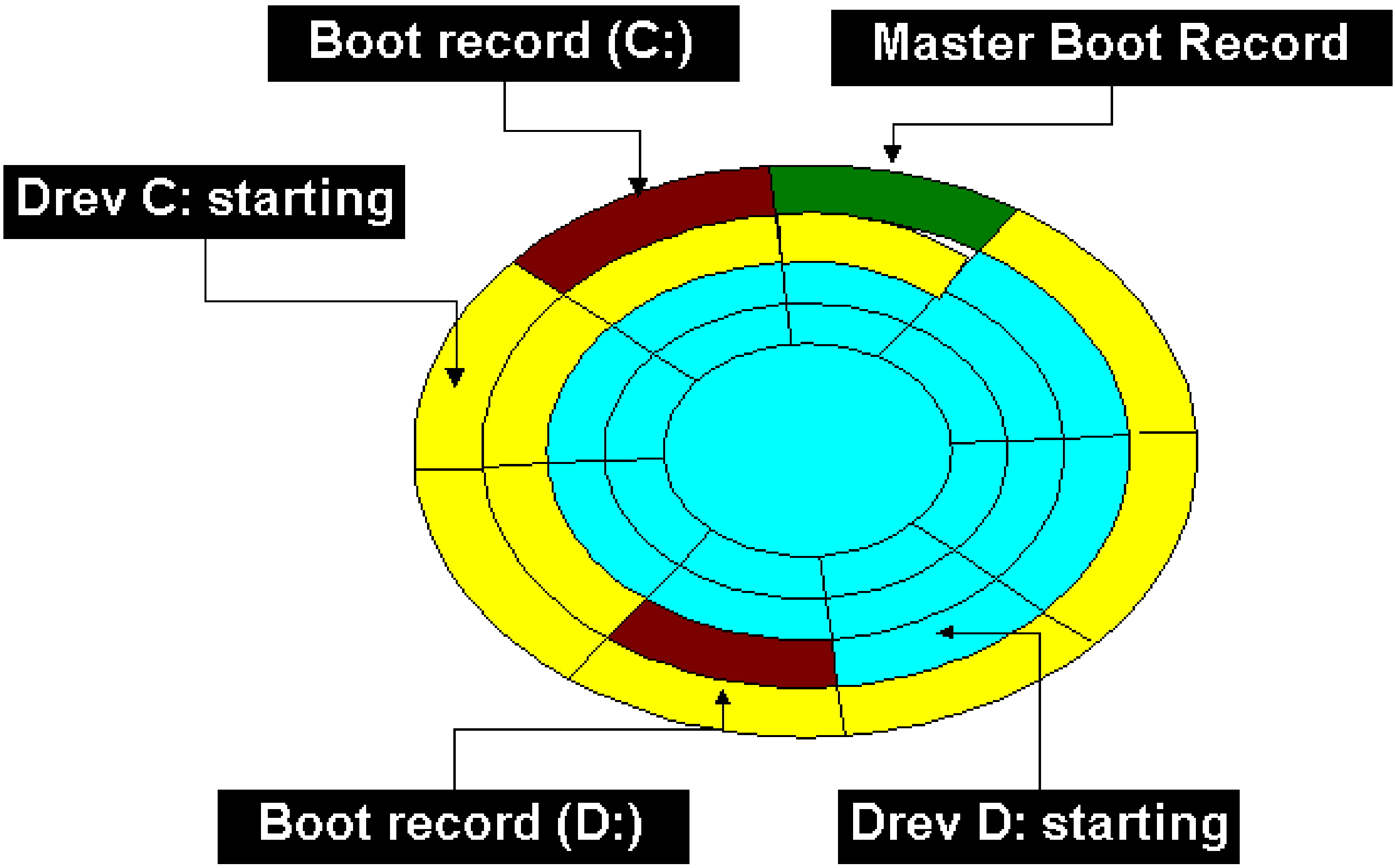
Therefore it is often seen that the hard disk is not utilized 100%. Look at this picture of FDISK, which has partitioned a 2 GB hard disk. The illustration is in Danish, but you'll se the same in English:

You can clearly see, that there are actually only two partitions. However, only the upper is assigned a drive letter (C:). The other partition consists of 43 MB unused hard disk, which FDISK identifies as Non DOS. It is not used, because you asked for a 2,020 MB partition. The remainder is left over. The 43 MB is not enough to bother to place in a new partition.

More boot records

When FDISK has partitioned the hard disk, the file system must be able to recognize this partitioning. Information about the location of beginning and end of each partition is stored in the first sector (number 0), which is called Master Boot Record (MBR). Then, regular boot records are stored in the beginning of each partition on the disk.

Here is a disk divided in two *logical* drives, which will be named C: and D:. The Master Boot record is in the first sector of the physical disk. It contains information about the two partitions. In the beginning of each partition we find a new boot record, describing that partition.



FDISK example

You use FDISK to divide the hard disk in one or more partitions. FDISK writes a MBR in sector zero. That divides the rest of the disk in logical drives, each of which is regarded as a "real" drive by the operating system. Let us look at the division of a large EIDE hard disk. It is sold as 5.1 GB. Actually, it holds 4.8 GB. Through FDISK, this capacity is distributed in three partitions. Here are the expressions, as used in Windows 95 version of FDISK:

- First a primary partition is created. We choose to assign it maximum size. That is 2,047 MB, corresponding to 2,146,467,840 bytes. That becomes our C drive, which is activated, so we can boot from there.
- We choose to establish an expanded DOS partition for the rest of the disk.
- The expanded DOS partition must be divided in logical DOS drives. We choose to make the first logical DOS drive the maximum allowable size. D drive will then be 2,047 MB, just like the primary partition is.
- A smaller part of the hard disk still remains. We will make that into a logical DOS drive. That will have 813,561,344 bytes, or 775 MB. That becomes the E drive.

Now FDISK reports that the disk has three drives. C: is the primary partition, D: and E: are two logical DOS drives, which are in the expanded partition. If we look at the physical hard disk, we find that it has a total of 9,974,720 sectors, 512 bytes each. After the partitioning, these almost 10 million sectors are distributed as shown below:

Physical sector number	Contents
0	Master Boot Record, which describes the entire hard disk
1 - 4,192,866	Drive C:

4,192,867 - 8,385,732	**Drive D:**
8,385,732 - 9,974,719	**Drive E:**

Note, that each of three drives has its own disk administration divided in boot record, FAT, root directory, and data area. If we select the C drive from above, we can see here how the sectors are distributed in the C drive partition:

Physical sector number	**Contents**
1	**Boot record**
2 - 513	**FAT 1 + 2**
514 - 545	**The root directory**
546 - 4,192,866	**Data area, which is divided in 32 KB clusters**

The primary partition and booting

There will always be one primary partition on the hard disk. Booting must be from the primary partition and the operating system is read from here.

The hidden system files

The core of the operating system is stored in the two hidden system files, which are always found in a primary DOS partition. In traditional MS-DOS, the files are named IO.SYS and MSDOS.SYS. These files have the same names in Windows 95, but the contents are changed slightly compared to the traditional DOS. This review is from the old fashioned DOS, but tells something general about the boot process of an operating system.

The DOS system formatted disk contains two hidden system files. The first, IO.SYS, *must* be the first entry in the root directory. MSDOS.SYS *must* be on entry number two.

Start-up on disk

When the start-up program has finished POST (Power On Self Test) and the loading of BIOS routines, the boot process starts. It follows the following steps:

- MBR is read first. The sector number of the primary partition's boot record is found here.
- A small boot program is read from the primary partition's boot record. That activates the loading of the two hidden files.
- IO.SYS is saved to working memory. Besides creating an interface "downwards" to the BIOS programs, IO.SYS includes a small program called SYSINIT. This program starts the next steps of the boot process.
- Now MSDOS.SYS is read from the disk.
- Then SYSINIT looks in root directory for a file named CONFIG.SYS. All commands in CONFIG.SYS are executed, gradually configuring the PC to have a ready and usable DOS.
- Then SYSINIT looks for the command interpreter COMMAND.COM. If that is not found, we will get an error message about this. When it is found, AUTOEXEC.BAT, which contains the last information for personal configuration of the PC, is executed.

That was a little bit about the boot process.

OS/2 Boot Manager

With OS/2's FDISK edition, you can divide the hard disk into more primary partitions. That allows use of the special Boot Manager, which comes with OS/2. Even if you do not use OS/2, you can still use Boot Manager let us say to have DOS /Windows 3.11 on one primary partition and Windows 95 on another. They will both appear as C drives, but you can only see one at a time. This, you control with the Boot Manager.

I hope you understand the importance of FDISK. It is a good program to be fluent in. Altogether, it is important to understand the file system, the boot process, etc.

There are two excellent utilities - Partition Magic and System Commander, which give further facilities to change the partitions and the start-up sequences, etc.

Long file names with VFAT in Windows 95

You can store long file names in Windows 95, which uses the VFAT file system. That is a 32 bit edition of FAT. VFAT was introduced with Windows 3.11, but the long file names did not become available until Windows 95.

The file systems in Unix, NT, and OS/2 have *always* been able to store long file names, but now Windows 95 can do it too. Also VFAT is compatible with regular FAT, which is smart. You can exchange files with other PC's - regardless of whether they can use long file names or not.

Actually, the VFAT file system is almost identical with regular FAT. But in a smart way Microsoft has been able to break the heavy 8.3 file name limitation, which limits regular FAT.

Physically, the file names are stored in a traditional 8.3 file name, which VFAT creates (without user control). The user can assign a long file name. As an example, a file is named "Ford Escort sales.doc". That will be translated to "FORDES~1.DOC", when the filename is registered by FAT.

The long file names may be up to 255 characters long, but they are translated to an "alias," which follows the traditional 8.3 FAT format. The trick is, that the long file name is written across multiple directory entries. Normally, one directory entry points towards one file, but in this case one file can occupy several root directories, each of which provides 32 bytes to the file name.

You should be happy about the long file names in Windows 95 - it makes it much easier to identify saved files. The only "danger" is, that you *must not* defragment the hard disk with a DOS based application. Then the long file names are destroyed. The files still exist, but you can only find them under their 8.3 name and that is an annoying experience, especially if you have thousands of files.

Last revision: 20 May 1998. To Overview. Total number of visits: since 10Dec96.

Click & Learn. **Module 6b. WWW.MKDATA.DK**

Windows 95 - operation and maintenance.

These pages describe hardware related experiences, which I have had with Windows 95.

Windows 95 and resources

Windows 95 is a very resource demanding operating system. That shows in the user applications: Office 95 and 97 are really big and heavy programs to work with. Then you might ask, why bother to use Windows 95?

The answer depends on your needs and ambitions. The office packages are undoubtedly the finest, most user friendly and most thoroughly planned office programs on the market - no question about that. They can work satisfactorily on your PC, but it requires some hardware. A lot of hardware indeed. Windows 95 has to run with a 5th generation CPU like a Pentium 200 MMX. The CPU is important, but RAM is even more important!

The need for RAM

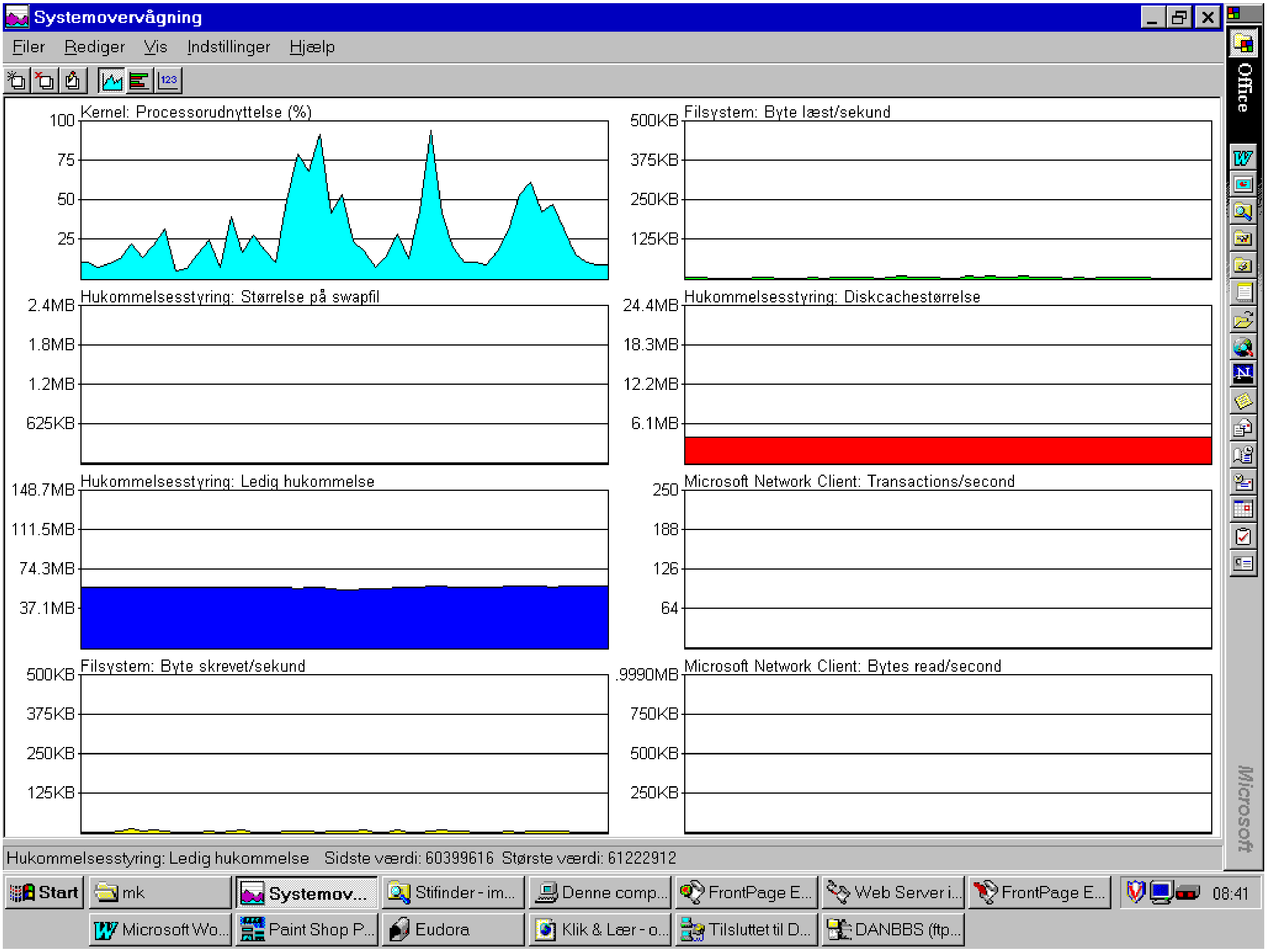
Windows 95 gobbles up memory. Therefore, sufficient memory is essential for its satisfactory performance. Try to check how much you really need - you will be surprised. The memory come from two locations:

- **The installed RAM**
- **The swap file, which is created automatically, when you run out of RAM.**

Windows 95 is clever in this way. It "extends" its RAM to the hard disk. If you only have 16 MB RAM in your PC, you can be assured that you have a sizable swap file on your disk.

The System Monitor

You can watch your use of memory with the excellent tool *resource meter*. You find it by going to: Start -> Programs -> Accessories -> System Tools -> System Monitor. You can add elements in the edit menu. That will allow you to see available memory and the swap file, as illustrated below (you may print this page, although the figure is in Danish):



You should check available *memory* and the size of the *swap file* over a period of time. Do this daily for a while and see how big the swap file gets. It is also a good idea to check the disk cache (more about that later), so that it does not occupy more than 2 MB. If the disk cache only occupies 1 or 2 MB, you can easily calculate your actual RAM usage by keeping track of *available memory* and the *size of the Swap file.*

The swap file

You need to keep an eye on the swap file, if possible eliminate it. It is a smart and necessary invention, but not very good in practice. Based on my conviction and experiences, many *breakdowns* in Windows 95 originate in swap file use.

If you get enough RAM in your computer, you can *eliminate* the need for a swap file. You really need to have 128 MB to be sure that this can be accomplished. I have personally worked with 64 MB and later 96 MB. However, I had to increase to 128 MB to eliminate these problems. Thus, it makes sense to spend the extra money for additional RAM, after you have used Resource Meter to determine your actual needs.

In some situations you may have to limit the swap file size to say 10 or 20 MB. Then Windows will tell you it has run out of memory when the swapping starts, and you can reboot. I prefer this to having a 300 MB swap file.

Disk cache

Another related problem is the built in disk cache in Windows 95. The cache is a portion of RAM, reserved for cache (buffer) for the hard disk. The disk cache is necessary, since it speeds up the hard disk a lot. However, it should not be bigger than 1 or 2 MB. The problem is that the disk cache really gobbles up RAM. It can easily eat up 20-25% of your RAM. An 8 or 10 MB cache is a total waste of RAM.

You can limit the size of your cache.

This is done by double clicking on the file System.Ini, which is found in C:\Windows. Scroll down until you reach the text [vcache]. Type in the two lines you see below and save the file. Do it soon. This is important!

- **[Non Windows App]**

  [vcache]
  MinFileCache=2048
  MaxFileCache=2048

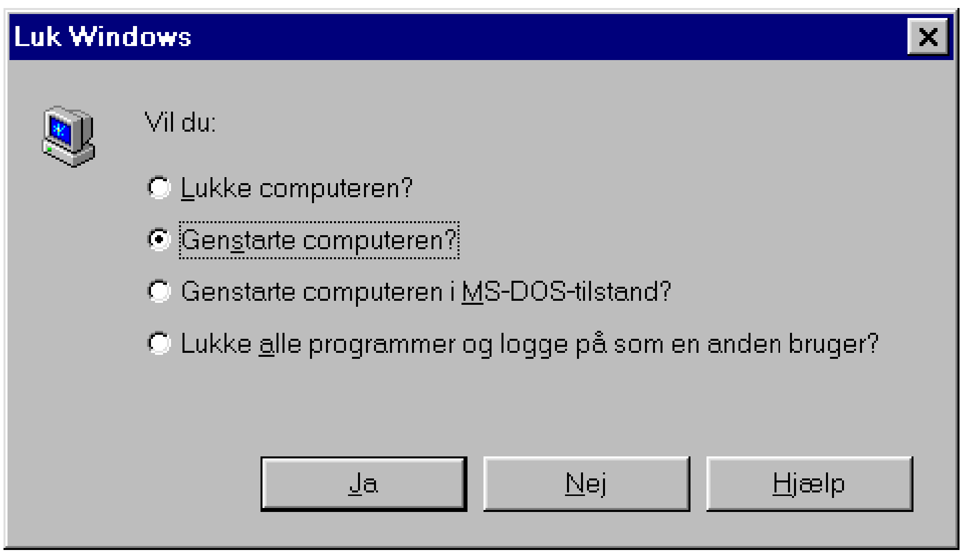
  [display]

The change takes effect when you restart Windows. I am convinced that 2 MB disk cache is sufficient - at least when you use the FAT16 file system. If you use Windows 95B with the new FAT32 file system, you should probably maintain 4 MB disk cache or maybe more. You see, Vcache has read a permanent copy of the FAT table. It occupies 2 MB in the new FAT32 file system. Read about file systems and about the cache copy of FAT.

Re-starting Windows 95

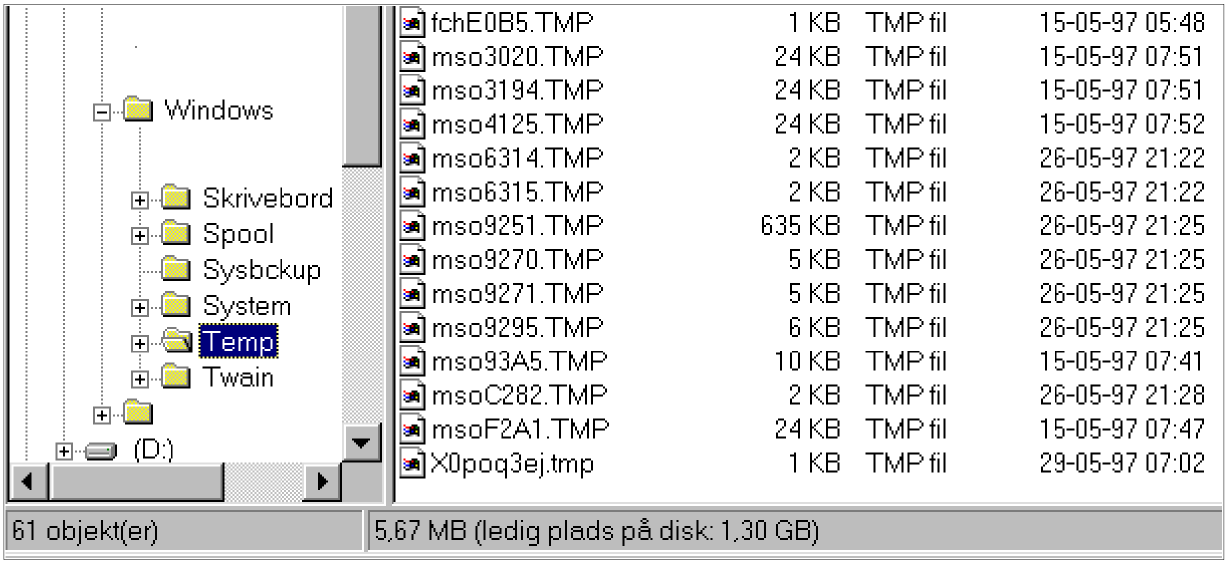
If you want to re-start the PC, you must click Start -> Shut Down.

If you want a faster re-start, you must depress the **[Shift] key**, while you click Yes (Danish figure):



Buckets of temporary files

Windows continuously creates temporary files. They are temporary files, which really need to be deleted. However, they are never deleted automatically - certainly not when Windows crashes, as it sometimes does. They are in the folder C:\Windows\Temp. You ought to check it routinely:



It is to your advantage to delete the temporary files. They just take up space and there can be hundreds of them. However, the problem is that you may not be able to delete all temporary files while Windows is running - some of them may be active. Therefore I recommend this simple method: put a line in your Autoexec.bat! You can find the file Autoexec.Bat (Autoexec) through Start -> Find. Right click on it and choose edit. Then type the line shown below and save the file:

- **echo Y | del c:\windows\temp*.***

This will cause all temporary files to be deleted at any start-up. The echo command adds a "Y" into the del command, so you do not need to confirm with a "Y" to execute the delete.

Also read in module 7a and module 7b about installation of monitor and video card in Windows 95!

Last revised: 20 May 1998. To Overview.

About operating systems and driver programs

Click & Learn deals primarily with hardware. In these pages I will cover the operating system as it connects downward towards hardware. The operating system is closely associated with the ROM-BIOS program routines, which are described in module 2a. The two program layers (operating system + BIOS) are called *system software* and it is very useful to understand their importance for the PC.

Let us start by studying what an operating system really is.

What is an operating system?

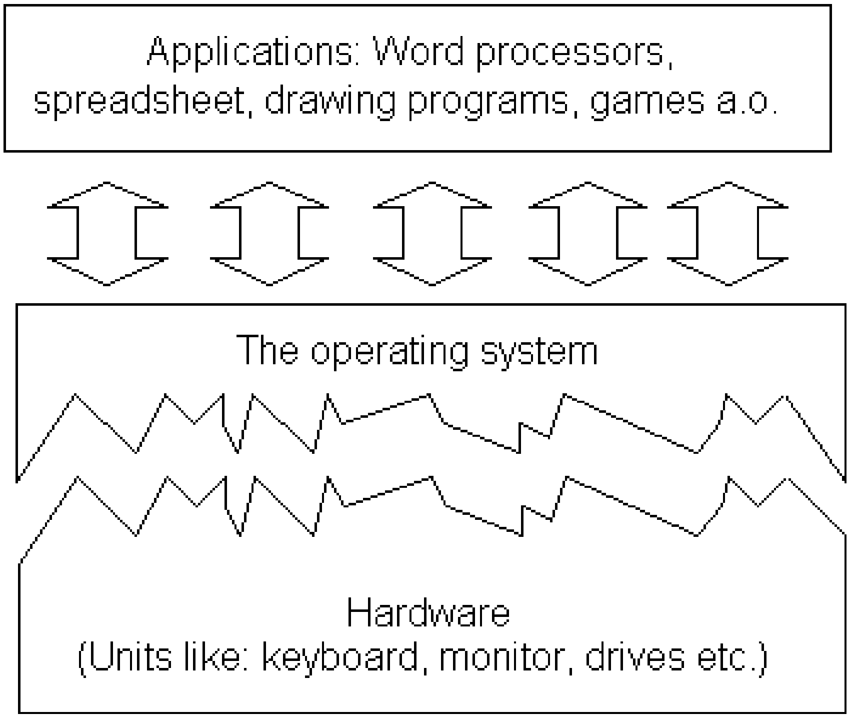
The operating system can be evaluated from different angles:

- An operating system is a number of files, which are read from the hard disk at the end of of the PC start-up routine.
- An operating system is a program layer. It connects to the the PC hardware, to facilitate optimal execution of the user programs.

The first definition does not say much. Let us start with the second: The operating systems links software and hardware together. It has to enable user programs, like Works, Office, etc., to function with all possible hardware configurations. You can imagine the relationship between hardware and user programs thus:

- Hardware is clumsy and dissimilar. There are untold variations of PC's. They can have one or another type hard disk, CPU, video card, etc. All of these various PC configurations behave each in their own way.
- The use programs are 100% similar. They are off the shelf products, which expect the PC to respond in a certain manner.

How do we make these two layers work together? Can we eliminate out the differences in the PC hardware, so a standard product like Works just functions? Yes we can. We read in an *operating system* - a system layer, which smoothes out and standardizes the hardware:

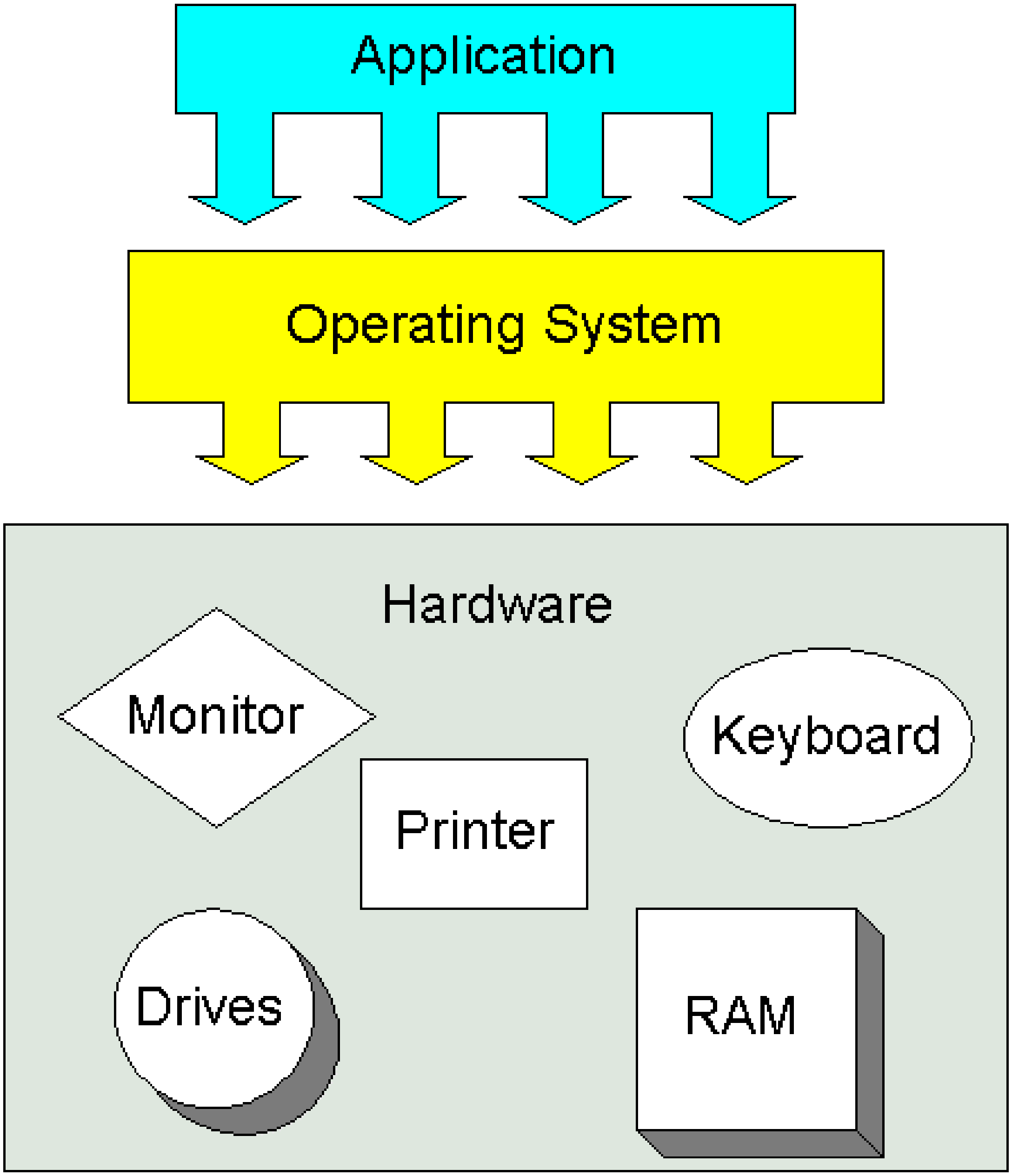


You should understand the operating system as a necessary layer, which smoothes out bumps and pot holes in your PC's hardware. This will give the user programs a stable, even work platform.

The operating system recognizes hardware

The PC's hardware represents *resources* relative to the user program.

Think of your word processing program: You want to print your text. The program issues a print order, expecting that the document will be printed as designed. The word processing program dispatches data according to your commands. How they are translated to signals understood by *your* printer - that is not the word processing programs problem. The printer is a resource relative to the word processing program. The connections to these resources is via the operating system. This holds true for all the resources, which are included in the PC hardware:



As you can see, the operating system has a very central function in the PC. So with that placement, it must be able to recognize all forms and types of hardware. There is no point in connecting a new mouse, if it does not work! Then what makes it work - the operating system. The system must recognize your mouse!

System software

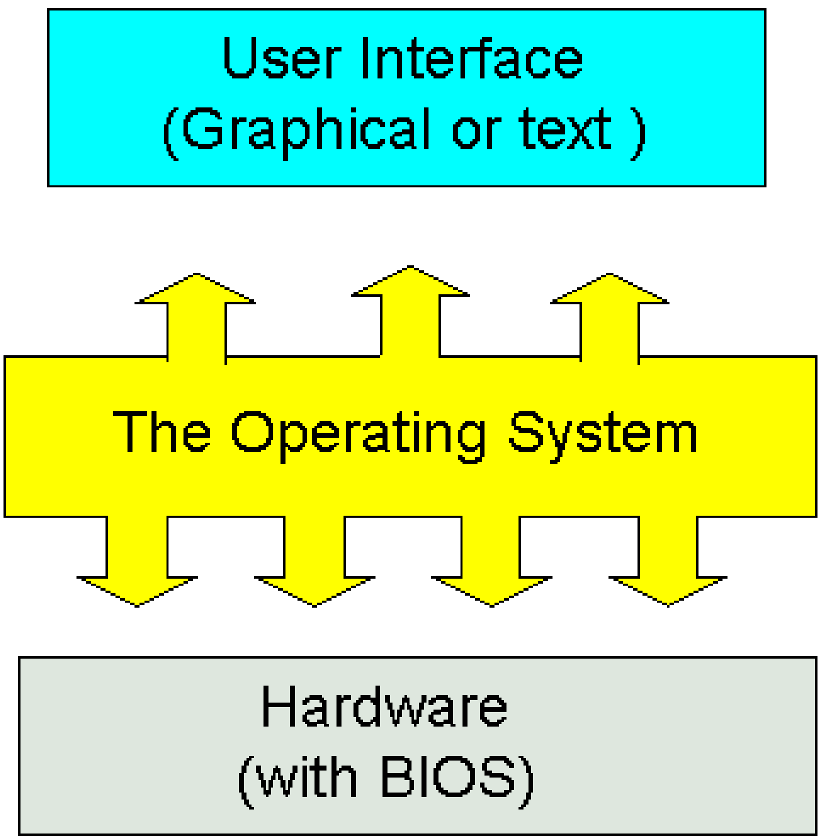
Together, the operating system and the ROM BIOS program routines form the layer on which the user programs "rest." When the PC has to work, an operating system has to be read from a disk. There are many different operating systems to choose from. However, the BIOS is always placed firmly and centrally in the PC hardware.

BIOS - firmware

One of the fundamental techniques in the PC design is the BIOS program layer. BIOS (Basic Input Output System) is a group of small programs, furnished by the PC manufacturer - also called *firmware*.

The BIOS routines are placed in the hardware - in a ROM chip - and are always available. Being stored in the hardware, they are functional regardless of which operating system they have to work with. So, in designing an operating system, one must pay close attention to the BIOS. The operating system must be able to work closely with the BIOS.

BIOS contains some very basic program routines, which handles data transfer between different hardware components. During PC start-up, the BIOS programs are the only accessible software. Later in the start-up process, the operating system is read,. It will then take control of the PC. The operating system has to provide a *user interface*, on which the use programs can rest. Thus, the operating system has two "faces": One pointing up towards the user and his/hers programs and one pointing down towards the system and hardware:



As computers have become more and more powerful, the user interface has become more graphic and user friendly. In a few years we will be able to address our commands directly to the operating system (you can do it already today with IBM's OS/2). Thus, the "upwards" face of the operating system will change greatly - supported by technological development. The "downwards" face - the operating systems interface with hardware - will change less. At least, the fundamental principles are the same as in the childhood of the PC.

BIOS or drive programs

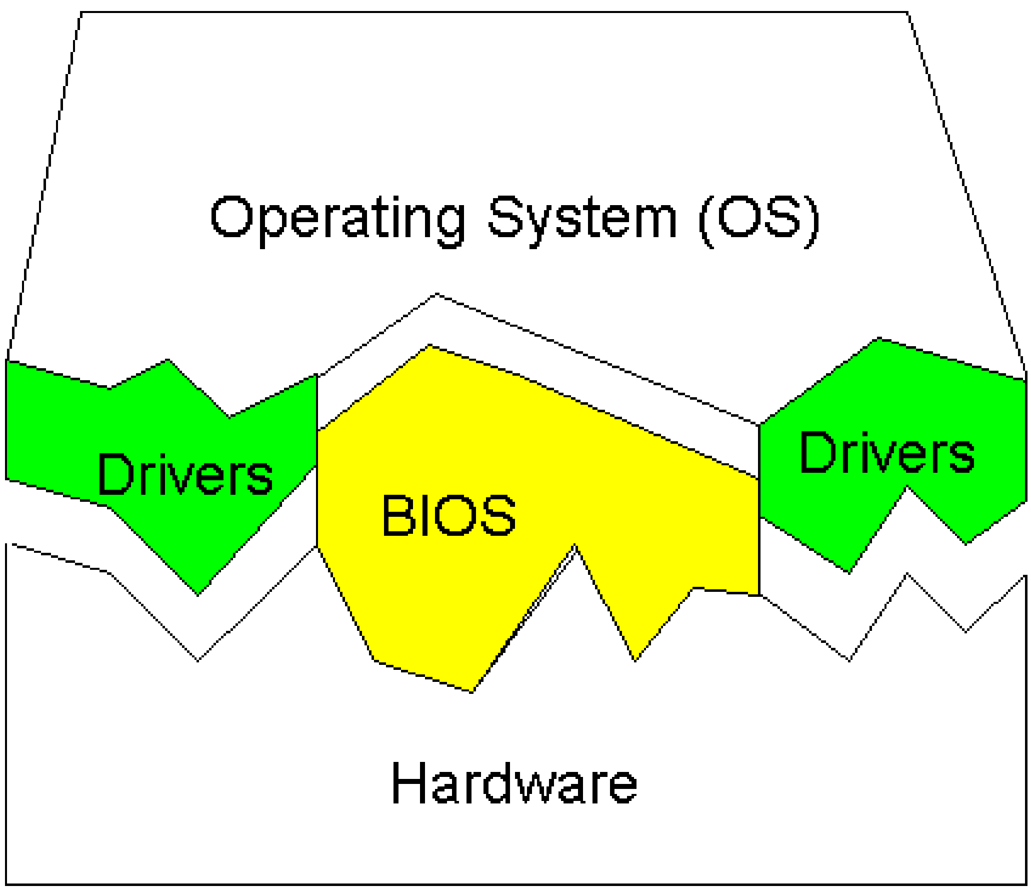
The operating system must be able to communicate with hardware. As we are going to see, this can be done in two ways:

- The operating system communicates directly with hardware through *drive programs.*
- The operating system utilizes the BIOS programs.

While BIOS is hardware specific program code, stored in hardware, the drive programs are small hardware specific program elements read from the disk together with the operating system.

Depending on which operating system is installed, both principles are used in various degrees. Since the BIOS programs consist of 16 bit code, it is typically DOS (a 16 bit operating system) which utilizes BIOS to a large degree. In the newer 32 bit operating systems, it is not efficient to use BIOS any more than necessary.

Here is a model, which shows the operating system with BIOS and drive programs (usually just called *drivers*):



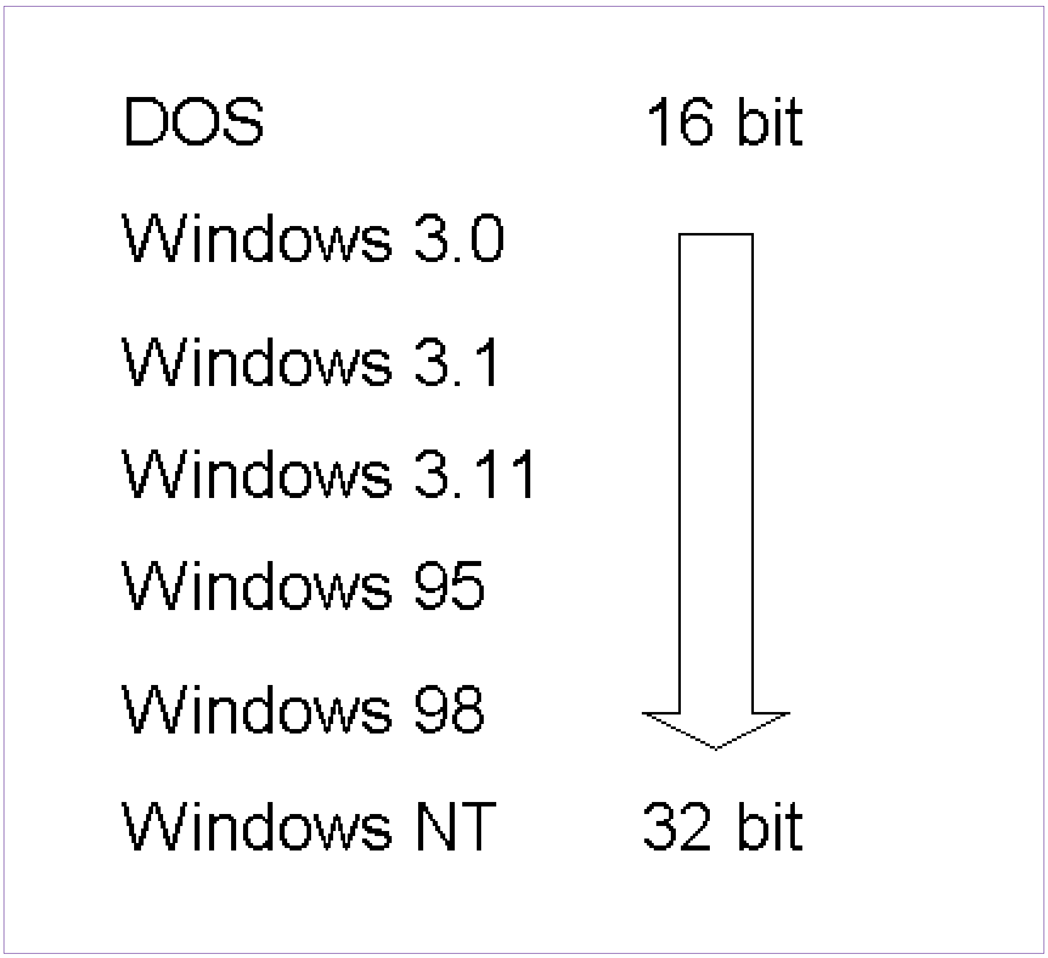
As you can see, the driver/BIOS functions are closely associated with the operating system. So let us look at that:

Which operating systems?

The operating systems have undergone a tremendous development since 1981. It all started with DOS, which was a 16 bit modification of the 8 bit operating system CP/M. DOS has then been further developed though the years. Since around 1990 Windows came into the picture. Windows started as a GUI (*Graphic User Interface*) for DOS . The PC booted with DOS. Then you could choose, if you wanted mouse and graphics on the screen with Windows. It was a supplement to DOS.

The Graphic User Interface (GUI) allows you work with a mouse instead of writing long command lines like copy c:\texts*.* d:\textbak\ *.* /s/v/, which is the standard in text based operating systems (like DOS).

DOS was designed for 16 bit computers, which the first PC's were. With Intel's 80386 the 32 bit technology was knocking at the door. Modern PC's are designed for straight 32 bit program execution. So we have seen a gradual trend in the PC operating systems from 16 bit towards 32 bit and this affects hardware design.



Only with Windows 95 did Windows partially separate from DOS. Today OS/2 and Windows NT are the most common pure 32 bit environments. Windows 95 and the coming Windows 98 are mostly 32 bit, but with some 16 bit remnants.

Protected mode

The 32 bit programs work in protected RAM sectors, with the CPU in *protected mode*. This allows the PC to multitask - more than one program can run concurrently and independently. That is not possible in 16 bit operating systems, where the CPU works in *real mode*.

A brief comparison of 16 bit and 32 bit operating systems can look like this:

Operating system	DOS	32 bit operating system (NT, OS/2, UNIX)
Users	Single user	Multiple users
Program execution	16 bit single task in real mode	32 bit multitask in protected mode
Screen appearance	Text based (poor quality graphics)	GUI - graphic interface with high resolution graphics
Hardware handling	Primarily BIOS	Custom designed 32 bit drivers for each hardware component.

DOS control of hardware

DOS is quite simple to describe, since it principally consists of only 4 parts:

- A boot record, which activates the operating system.
- The file IO.SYS, which is interfaced to ROM-BIOS with installation of *device drivers*.
- The file MSDOS.SYS. That is the core of DOS, handling the file system and program execution.
- The file COMMAND.COM, which provides the command line, the text based user interface.

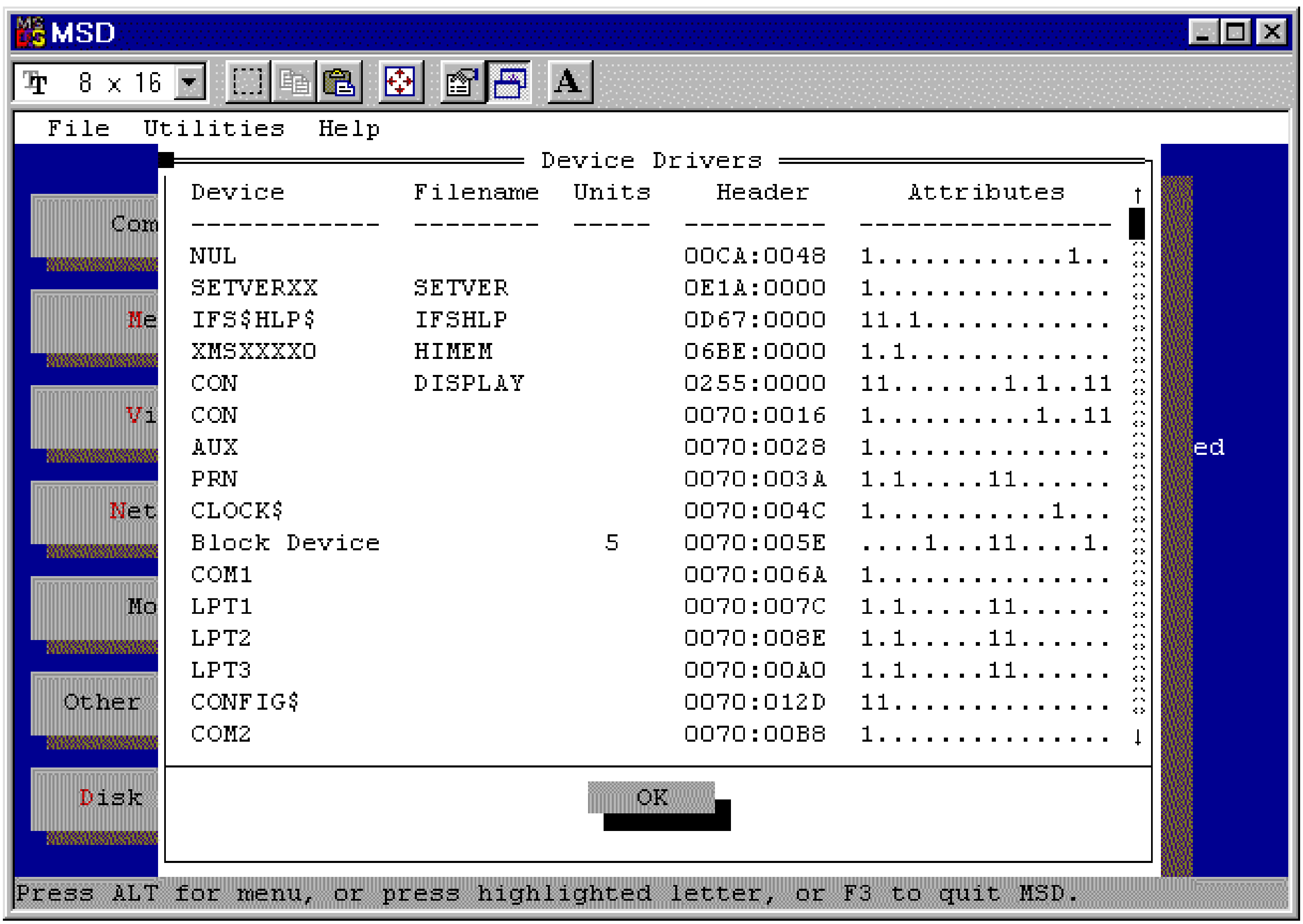
When we talk about hardware control, it is done through IO.SYS. That is a program which reads the ROM BIOS code and converts it to DOS's own device drivers.

The smart thing about DOS is that the operating system can be *expanded* with external device drivers. IO.SYS reads them via the start-up file CONFIG.SYS. First device drivers are read from ROM BIOS. Then any possible additional drivers are read from disk. In that way DOS can handle hardware units which did not exist when the PC was originally configured.

A final option to handle hardware from DOS programs is to write special drivers for the individual *user program*. Many DOS games come with their own graphics drivers (they have to recognize all graphics standards on the market!). Another classic example is the word processing program WordPerfect, which in its prime (version 5.1) came with drivers to more than 500 different printers!

Unit	Example of DOS device drivers
Hard disk	BIOS
Video card	BIOS
Mouse	MOUSE.SYS
CD ROM	ATAPI.SYS + MSCDEX.EXE
Printer	Internal drivers in the user program (like WordPerfect 5.1)

The device drivers can be seen with the program MSD. Here is a picture from my Windows 95, where you can clearly see the names of the device drivers (CON, PRN, LPT1 etc.):



All these device drivers are in 16 bit program code.

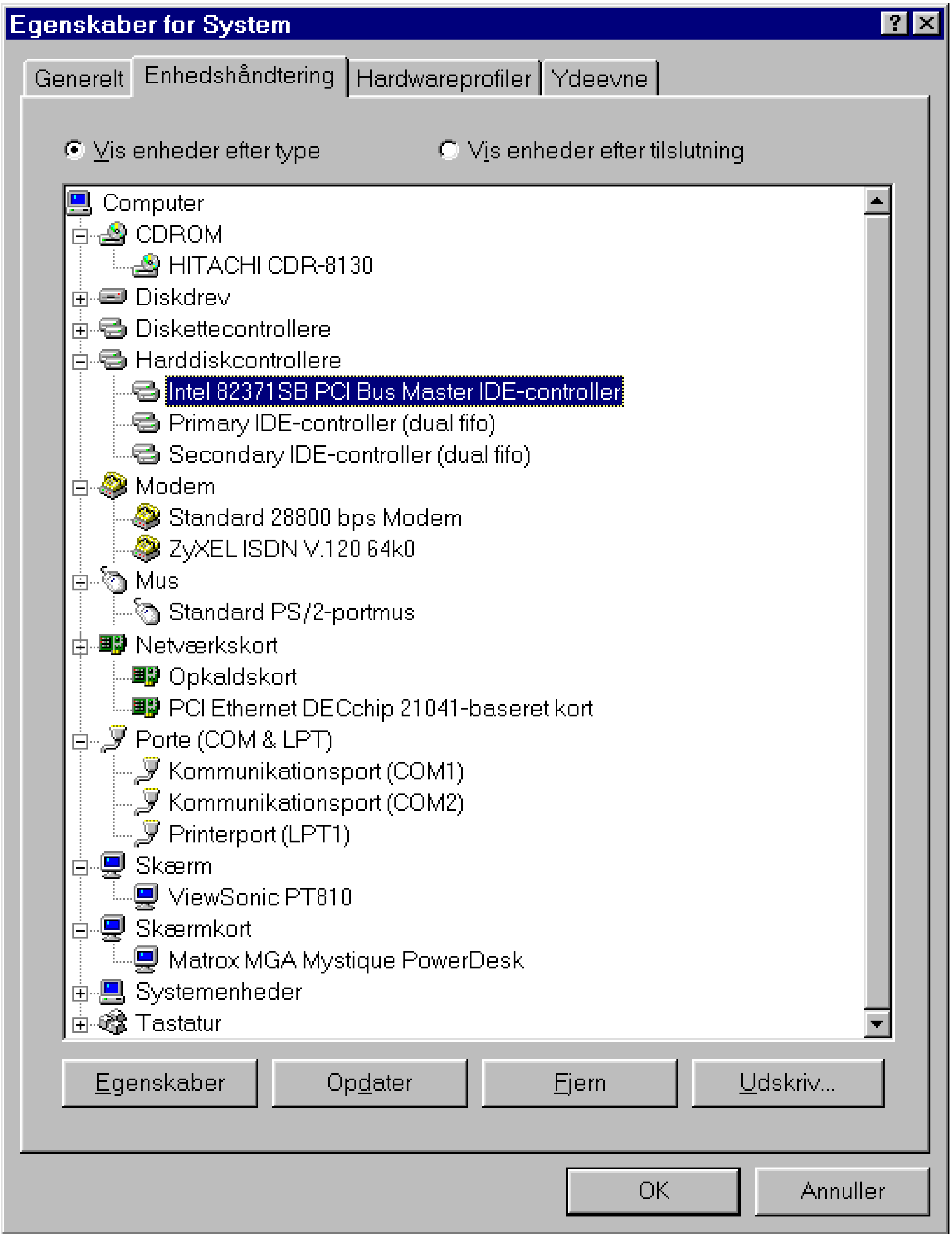
32 bit drivers

In 32 bit operating systems, you use 32 bit drivers instead of ROM BIOS. This means that software suppliers like Microsoft and IBM must be able to supply drivers to all conceivable hardware. The advantage is, that once the operating system has installed drivers, all user programs operate alike relative to hardware.

It is en enormous project to supply drivers. Especially OS/2 has suffered problems in getting the right drivers on the market. For many years, IBM for example did not supply OS/2 drivers for Canon printers. That was part of my reason to drop that operating system. Regarding driver installations, Windows 95 is unquestionably the best operating system.

Windows 95 supports plug and play. The operating system searches the PC for hardware. Often all drivers (to CD-ROM, network controller, sound card, etc.) are installed automatically. The drivers can be seen under System in the control panel.

Let us look at my common EIDE hard disk. The hard disk operation is regulated by an EIDE controller on the system board. Therefore, Windows 95 must have a driver installed to this controller. We can find it easy. Go to: My computer -> Control panel -> System -> Computer. Then click input/out [I/O]. Here we see a number of hardware units. An 82371SB PCI Bus Master IDE controller, which regulates the hard disk is highlighted:



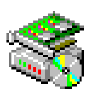
Actually, you can see a long list of drivers in the picture above. Windows 95 has installed most of them during Windows installation, but I have added some. That includes the ISDN modem, Logitech MouseMan (which you don't see) and the video card (Matrox). I always have these drivers on the hard disk (in the folder C:\Utils\Drivers). That makes it easy to install them after a unforeseen but necessary re-installation of Windows 95.

The quality of the drivers is very important. The drivers are extremely important for video cards. You often hear that a new driver has been developed for this video card and it improves performance by 40%. Then rush to download it (from the manufacturers Internet server) and install it. Don't forget to save it on disk for future use!

You also have to be cognizant about the system board chip set. Often Windows 95 installs a good standard driver, but new chaplets may contain facilities which require a new driver. That can be found on a diskette, which comes with the system board.

Installation of new drivers

You install new drivers in Windows 95 with "add new hardware" found in in my computer -> control panel:.



Don't let Windows 95 search for hardware. Instead choose yourself. Then you have to select the particular hardware from the list and in the next screen click "Have diskette... " Learn this technique if you experiment with your PC and want maximal benefits from your hardware.

Read also in module 7a and 7b about installation of monitor and video card in Windows 95!
And module 7c about DirectX.

Click & Learn. Module 7a. WWW.MKDATA.DK

About the video system

The video system (of which the monitor is a part) is one of the most important components in the PC. It affects directly your pleasure of working, and actually also your health. At the same time, the video system shows the biggest variation between different PC's. Read my coverage of this subject here.

Introduction

All computers are connected to some type of *display*. That is called a monitor . Monitors are available in many different types and sizes (generally 12 to 21 inches diagonal screen size). The monitor is a part of the computer video system. To understand how to obtain a good screen image, we need to look at the complete video system. It includes three elements:

- The graphics card (also called the video card or video adapter). It is an expansion card, which generates electric signals to the monitor.
- The monitor itself, which is connected by a cable to the video card.
- A device driver which Windows 95 uses to control the video card, to make it send the correct signals to the monitor.

These three elements must be fitted and matched to achieve quality images. Even the finest and most expensive monitor will only render mediocre images if it is connected through a low quality video card. All video cards depend on the right driver and proper settings to function properly – otherwise the card will not perform well.

In these pages, I will review the complete video system. First you can read about the video image construction, pixels. resolution, and refresh rate. Those are very central subjects. Later, we will look at different monitor and video card types. Finally, put it all together in Windows 95.

Fast development

The video system has developed as explosively as the rest of the PC during the last 10 years. These improvements have occurred in different areas:

- The monitors – both the tubes and the electronics continue to improve. They render better and better images - sharper, with better resolution and better colors.

- New monitor types are arriving on the market. The flat TFT screens are on their way to the stores. In a few years, they will become the new screen standard.
- The video cards are getting faster. They can deliver better images, which the new monitors are capable of producing. The user gets more tuning options. New RAM types and the AGP bus will increase speed. Video presentation and 3D games are other areas of development, which will change the video card standards.

The video system is a sub system in the PC, with its own technological development. At the same time, monitors and video cards are areas, where manufacturers and dealers often cut corners. As an ordinary user, you can improve your screen images significantly with careful planning. That holds true when you buy your PC - you must select your video system carefully. It also holds true for existing video systems, where better drivers and software optimizing can help produce the optimal screen image. We will look at that in these pages.

Concepts and terminology

When we talk about screens, there are currently three different types to choose from:

- CRT – (*Cathode Ray Tube*) the common type screens. They are found in different technologies, such as *Invar* and *Trinitron*.
- LCD – *(Liquid Crystal Display)* flat and soft displays. TFT – is the most expensive display of this type. They are found in better grade laptops, and are available also for use with desk top PC's. Their current price is in excess of $2,000 for a 15,5" TFT screen. The TFT screen is also called a "soft" screen, since the images appear softer than from Cathode Ray Tubes..

Common principles

The principles in these screen types are quite different, but the screen image design rests on the same concepts:

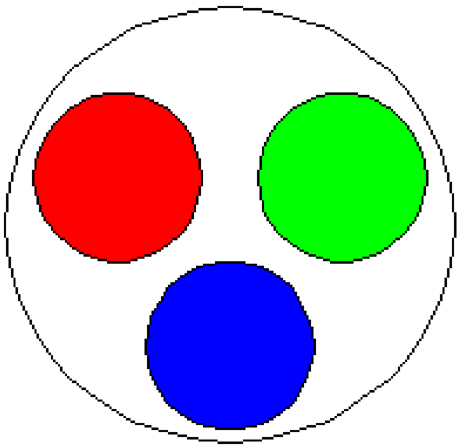
- Pixels. The screen image is made of pixels (tiny dots), which are arranged in rows across the screen. A screen image consists of between 480,000 and 1,920,000 pixels.
- Refresh rate. The screen image is "refreshed" many times per second. Refresh rates are measured in Hertz (HZ), which means "times per second".
- Color depth. Each pixel can display a number of different colors. The number of colors, which can be displayed, is called color depth. Color depth is measured in bits.
- Video RAM. All video cards have some RAM. How much depends on the desired color depth. Video cards usually have 1, 2 or 4 MB RAM for normal usage.

These concepts are central to the understanding of the video system. Since the CRT screens are still by far the most common, they will form the basis for this review.

Pixels – the basic element in the screen image

When you look at a screen image, it actually consists of thousands of tiny dots. If you look close you can spot them. Each of these dots is called a *pixel*. That is a contraction of the term Picture Elements.

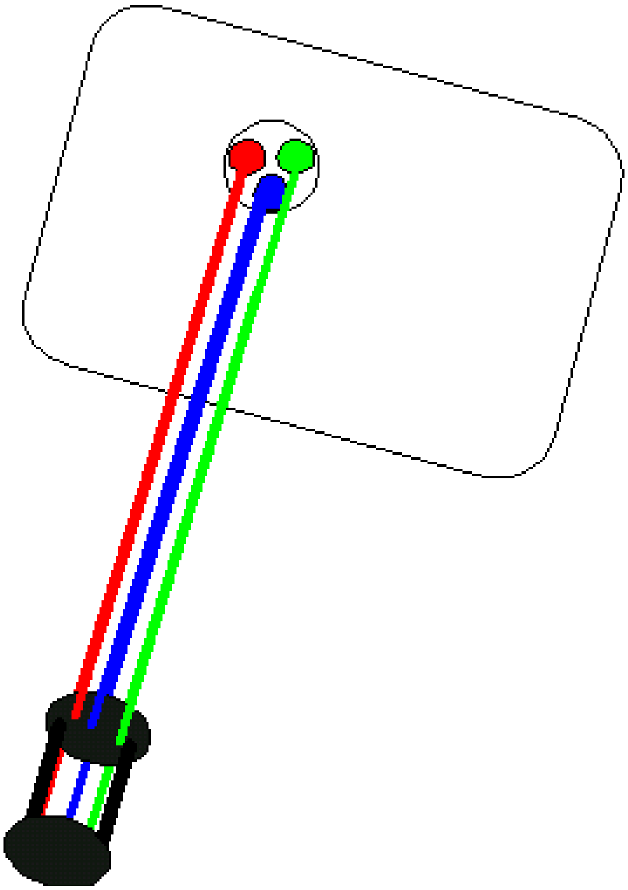
In an ordinary screen, each pixel consists of *three* colors: Red, green and blue. Thus, there are actually three "sub dots" in each pixel. But they are so small that they "melt" together as one dot:



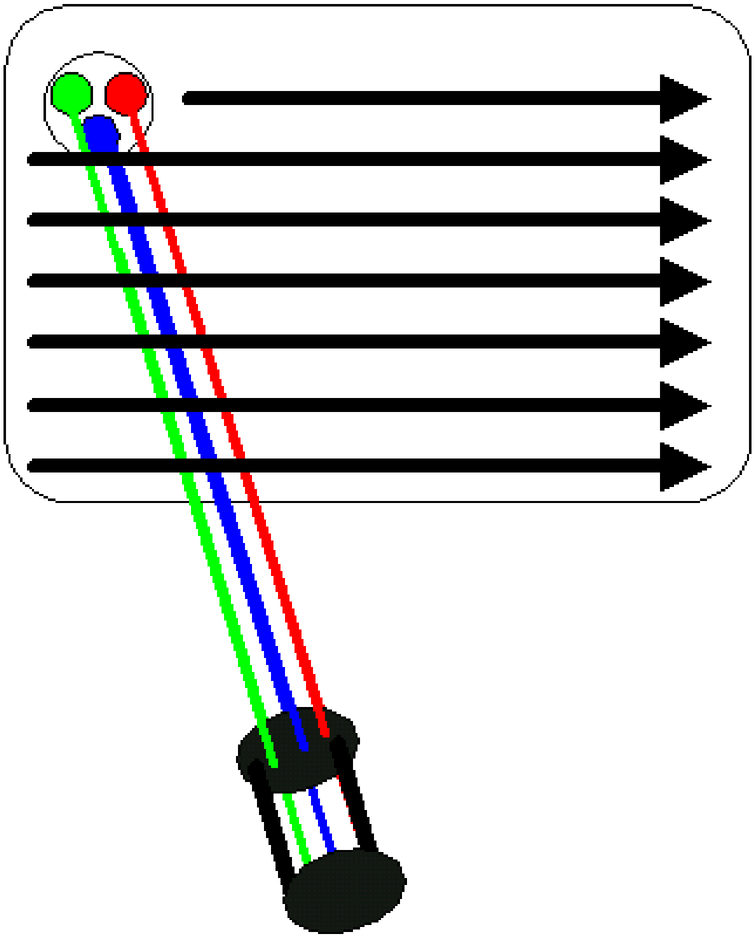
The individual pixel or dot then consists of three mini dots, also called *trio dot*. Some screens do not have round dots, but they work the same way. With the three basic colors, each of which can be assigned with varying intensity, you can create many different colors.

The cathode ray tube

A traditional picture tube is like a big glass bottle. There are three electron guns in the narrow end. They fire towards the large flat surface facing the user.

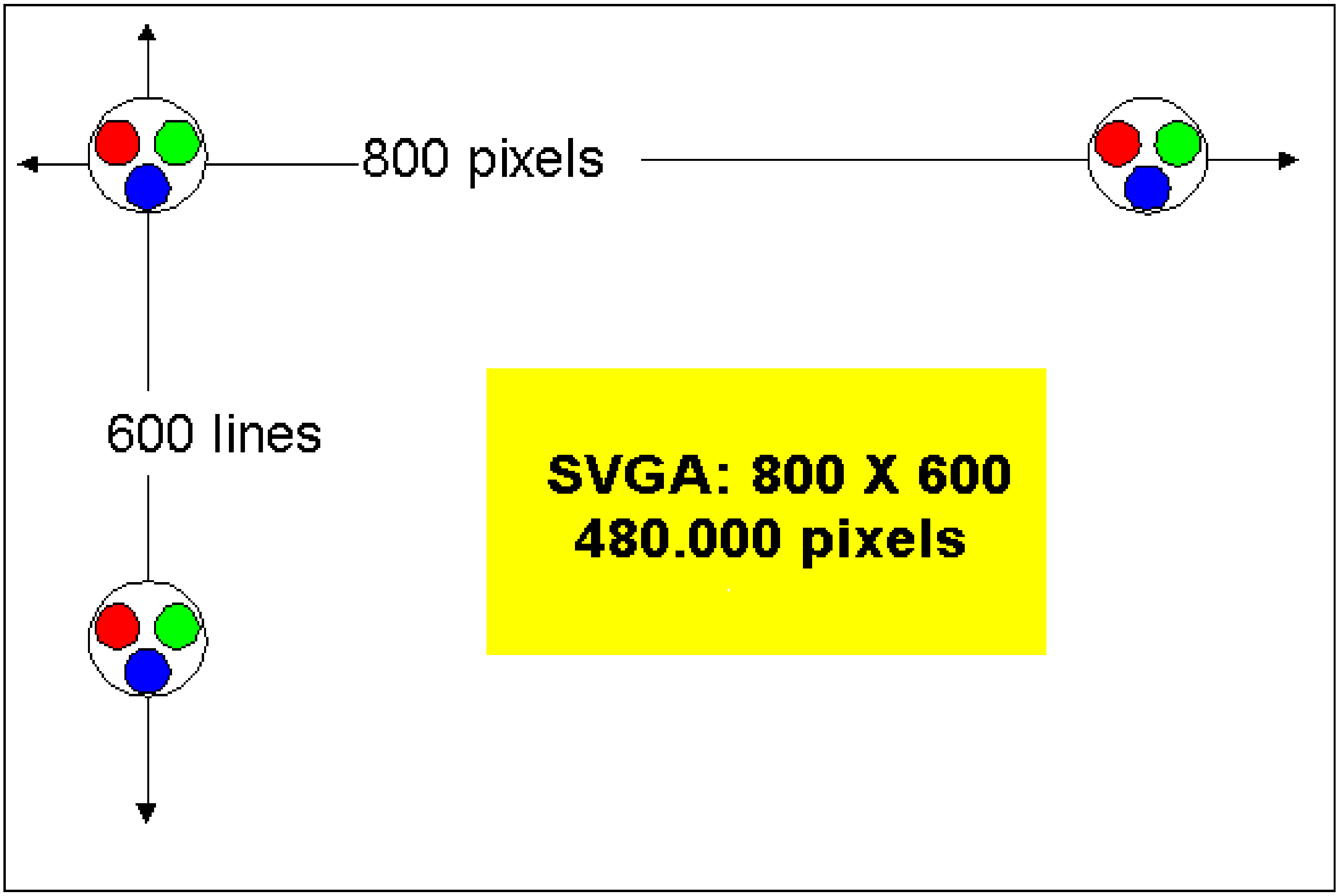


The inside of the glass surface we look at is coated with tiny phosphorous dots. They are arranged in groups of three – a red, a green and a blue phosphorous dot. Together they make a pixel. These dots light up, when hit by elections from the electron gun. Each of the mini dots is hit by its own electron gun.
The more powerful the beam is, the brighter they get. The electron beams are guided by electromagnets, which bend the beams, so they hit the exact desired phosphorus dot.



The electron beams sweep across the screen very fast. Each of the three electron guns must scan its intended color mini dots continually, from left to right, line by line from top to bottom, typically about 70 to 85 times per second. The *beam intensity* can be adjusted for every mini dot, to adjust the color.

A typical screen image could consist of 480.000 pixels. That is called a 800 x 600 image. There are 800 dots in each horizontal line, and there are 600 lines from top to bottom of the screen. That adds up to 480,000 pixels.



Greater resolutions

The greater the number of pixels in the screen image, the better the *resolution*. And the greater the resolution, the sharper the image appears.

The lowest resolution seen in modern PC's is found in text based DOS screen images, which are 640 x 480 pixels. That is called a VGA image. VGA was the standard, until Windows came on the market. Back in the eighties, there were even lower standards, like CGA, which I will not even describe.

As the PC's got more powerful, around 1990 a demand developed for better screen resolutions. Windows is a graphic environment, and it works fine in all screen resolutions. The same programs work as well in 640 x 480 as in higher resolutions. Many DOS games also demanded better screen quality. Anyway, VGA was the last "real standard" working on any PC. Screen resolution was since improved relative to VGA, and the term SVGA (Super VGA) came into use. Later came XGA and other names, which each described different resolutions.

Actually, the terms SVGA and XGA are not used much anymore. Instead we are looking at resolution, image frequency and color depth. But, let us stay with the resolution. It ties in with screen size, the bigger the screen the bigger the possible resolution. Below, you see a table with different resolutions:

Standard	Resolution	Number of pixels	recommended screen size
VGA	640 x 480	307,200	14"
SVGA	800 x 600	480,000	15", 17"
SVGA	1024 x 728	786,432	17", 19"
XGA	1152 x 864	995,328	17", 19", 21"
XGA	1280 x 1024	1,310,720	19", 21"
XGA	1600 x 1200	1,920,000	21"

Screen and resolution need to be matched

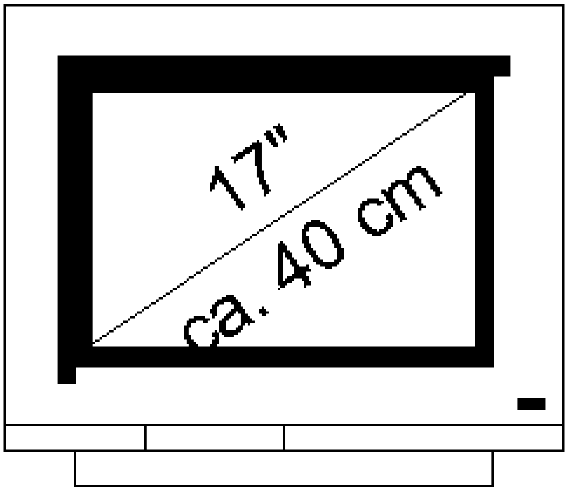
The greater the resolution the more detail you can view on the screen. On a Windows 95 desk top, you can see 2-3 as many icons in a 1280 x 1024 resolution as in a 800 x 600.

The individual monitor can be set to different resolutions. However, not all resolutions are suitable. On a small screen the icons get too small, if you choose too high a resolution. Therefore, resolution and screen size must be matched!

You can not judge a monitor just by its resolution. Equally important is refresh rate and color depth. They will be described later. But first:

Screen size

Monitor screen sizes are measured in inches, just like TV sets. The most common sizes are 14", 15", 17", 19" and 21" screens. The measurement is the diagonal size of the screen



However, the nominal size of the common CRT screens does not give a true description of the visible size. The nominal size is the *internal* diagonal of the the picture tube. However, the *visible* diagonal is smaller!

The visible diagonal of a CRT screen is always about 10% smaller. Therefore, the visible image on a 17" CRT screen and a 15.5" LCD screen is about the same. Here is a comparison of different screen sizes:

Screen type	Typical visible diagonal
12.1" LCD	30.5 cm
14" CRT	31 - 33 cm
13.8" LCD	35.0 cm
15" CRT	32 - 35 cm
15.5" LCD	39 cm
17" CRT	39 - 41 cm
19" CRT	44 - 45 cm
21" CRT	49 - 51 cm

Colors

Colors are a must! Good PC’s must be able to display many colors on the screen.

Variable light intensity

The colors are created by varying the light intensity of the three basic colors (red, green and blue). Traditionally we work with the following degrees of *color depth:*

- 256 colors (8 bit color)
- 65,536 colors (16 bit color, also called *65K* or *HiColor)*
- 16 million colors (24 bit color, also called *True Color)*
- 4 billion colors (32 bit color, also *True Color)*

For ordinary users, 256 bit colors is sufficient to render high quality reproduction of photos. However, 256 bit colors have limited usage, the colors can get rather coarse. 24 and 32 bit colors should be the choice for graphics artists and professional photographers.

But how are these color variations created on the screen?

Three colors, each in their own depth

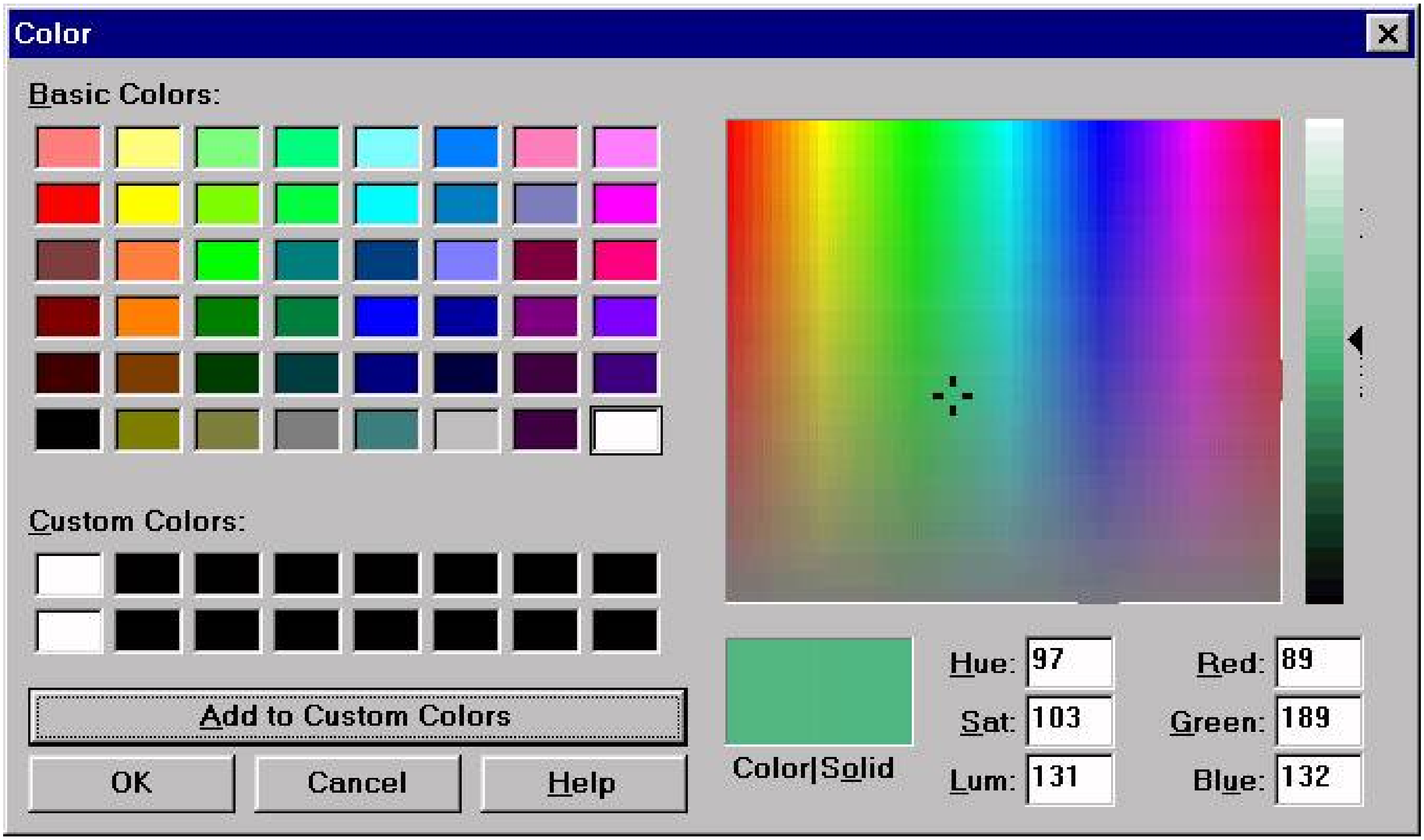
Each pixel is composed of three mini dots. The three mini dots are so close together, as to be viewed as one pixel. But three mini dots can be illuminated each with their own intensity. This allows us to produce many different colors.

Let us imagine to vary the light intensity of the three mini dots like this:

- Red in 1000 steps
- Blue in 1000 steps
- Green in 1000 steps

That will result in 1000 x 1000 x 1000 colors, a total of one billion different colors. In real life, this is not that dramatic and they are not counted in thousands. Instead, we are talking about so many *bit* colors.

Here you see color adjustment from the program Paint Shop Pro. There are 48 standard colors, but you can add and mix new colors as you wish:



---

16 or 24 bit colors

We must be able to vary the three basic colors in a number of steps. Typically. we use 8 bits for each color. 8 bits provide 256 possible variations – from 0, giving a light intensity of zero – and up to step 255, giving maximum intensity of that color. That will provide the following possible color variations:

- Red in 256 steps
- Blue in 256 steps
- Green in 256 steps

All together there are 3 colors, each of which can be controlled in 256 steps. To identify each of the 256 steps, we need one byte of data. However, we work with bits. Thus each color requires 8 bits of data.

Bear in mind that the PC must remember the entire screen image. Then how much data is required to remember one color out of 256 possible? That "costs" 8 bits. Each pixel has three colors, each of which must be adjustable in 256 steps. That "costs" 3 x 8 bits. Therefore, we refer to them as 24 bit colors.

24 bits can produce 256 x 256 x 256 colors = about 16 million different colors.

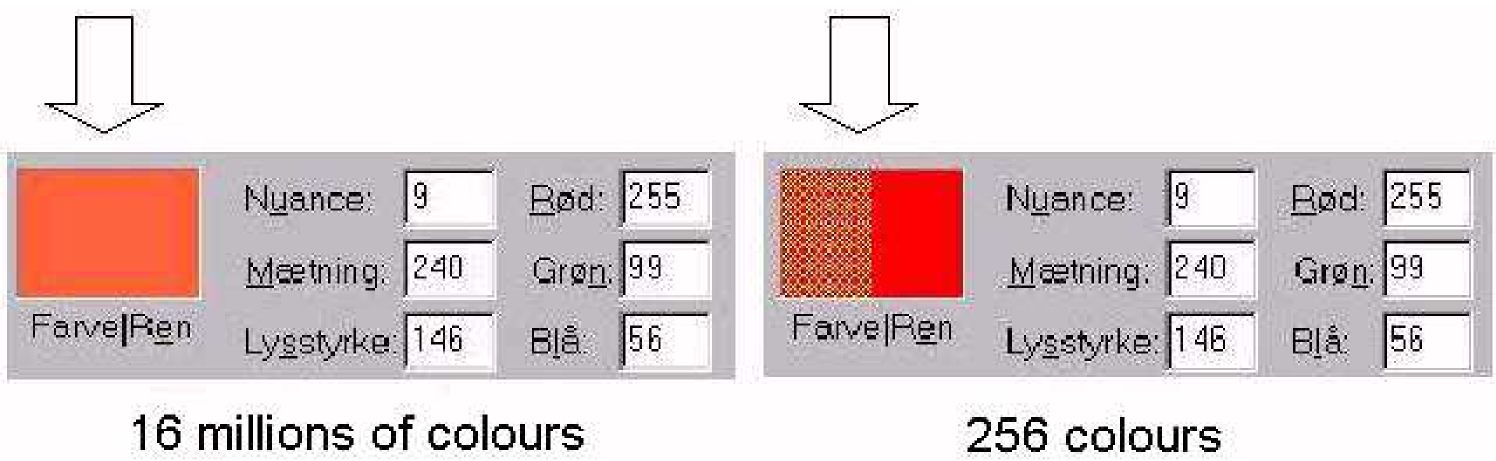
Since the PC works with bytes, in practice we use 8, 16, 24, or 32 bit colors. That produces a palette of colors, as we started to see. In a 16 bit color image, each pixel can show any one of approximately 16 million colors. It requires 16 bits of data to describe the color choice of each pixel.

A good video system can then reproduce many colors – at least 16 bit colors! The more pixels and the more color depth – the greater is the memory demand, the more bytes are required. We will return to that later.

An orange color – an example

Below you see how a given color is composed of red, green, and blue, each in their defined intensity. The scale goes from 0 to 255. Since the video system can show tons of colors, this color is seen as pure orange:

Here you see the same color setting on two different screens. On the left side there are 16 million available colors, giving a pure rendition of the color. On the right side the screen can only produce 8 bit colors. There are only 256 colors to choose from and this particular color cannot be reproduced properly:



The pure color (to the left) is available on PC's with 16 bit video system or better.

Refreshing the screen image

In traditional screens, the electron gun continually sends out very precisely aimed beams of electrons, moving from pixel to pixel. The beam actually flickers, as it sweeps the screen. Each dot on the screen receives a quick flash of electrons, before the beam moves on to the next dot. And the beam intensity is varied from dot to dot.

The phosphor coating on the screen has the peculiar ability to light up, when hit by electrons. But the light quickly fades away. In practice, The electron beam "visits" again, before there is any visible fading of the light. It looks to us as a steady picture on the screen, but actually it *flickers* every time the electron beam hits the phosphor coated dots.

The screen works overtime

Typically, each pixel is hit 60, 70, 75, or 80 times per second. Thus, the electron gun must move extremely fast to make 18 million or more hits per second. If the image is refreshed 75 times per second, we talk about a *refresh rate* of 75 HZ.

The video card issues the refresh signals, thus controlling the refresh rate.

Here we see a screen with a resolution of 1280 x 1024 and a refresh rate of 75 HZ. That requires the electron gun to make 98 million pixel hits per second! That screen works at a very hectic pace – which can sometimes result in beam contamination.

High refresh rate

The screen image appears more steady, the higher the refresh rate. You see the same in TV, where traditional sets have a refresh rate of only 50 Hz. Some manufacturers now produce TV sets with 100 HZ refresh rate. Some claim that they cannot notice the difference. However, once you have been used to 100 MHZ refresh rate, it is uncomfortable to return to 50 HZ. Similarly with PC monitors, only here we have more options.

Older and inferior screens can only work at 60 HZ, which produces a low quality, flickering image which is not suitable for Windows. The general consensus is that 70 HZ produces an acceptable image. I find 75 HZ acceptable, but 80 or 85 HZ is better when you have to work many hours daily in front of the screen.

Note: refresh rate is also called *vertical frequency* or *vertical refresh rate*, but I have chosen to use the term refresh-rate.

The higher the refresh rate, the better quality monitor you need. If you want both high resolution and high refresh rate, you will need both a high quality monitor and a high quality video card. The bigger the screen, the more it must be able to produce.

Screens can always run with higher refresh rates in lower resolutions. Here are three examples, showing how the screen performance drops with resolution.

Screen	800 x 600	1024 x 768	1280 x 1024	1600 x 1200
Standard 15"	75 HZ	70 HZ	60 HZ	-
15" Trinitron	90 HZ	80 HZ	75 HZ	-
17" Trinitron	110 HZ	100 HZ	90 HZ	85 HZ

For the screen to deliver images at the desired refresh rate, both screen and video card must be matched to the correct specifications - the higher your demands are, the higher the cost will be.

More about screens

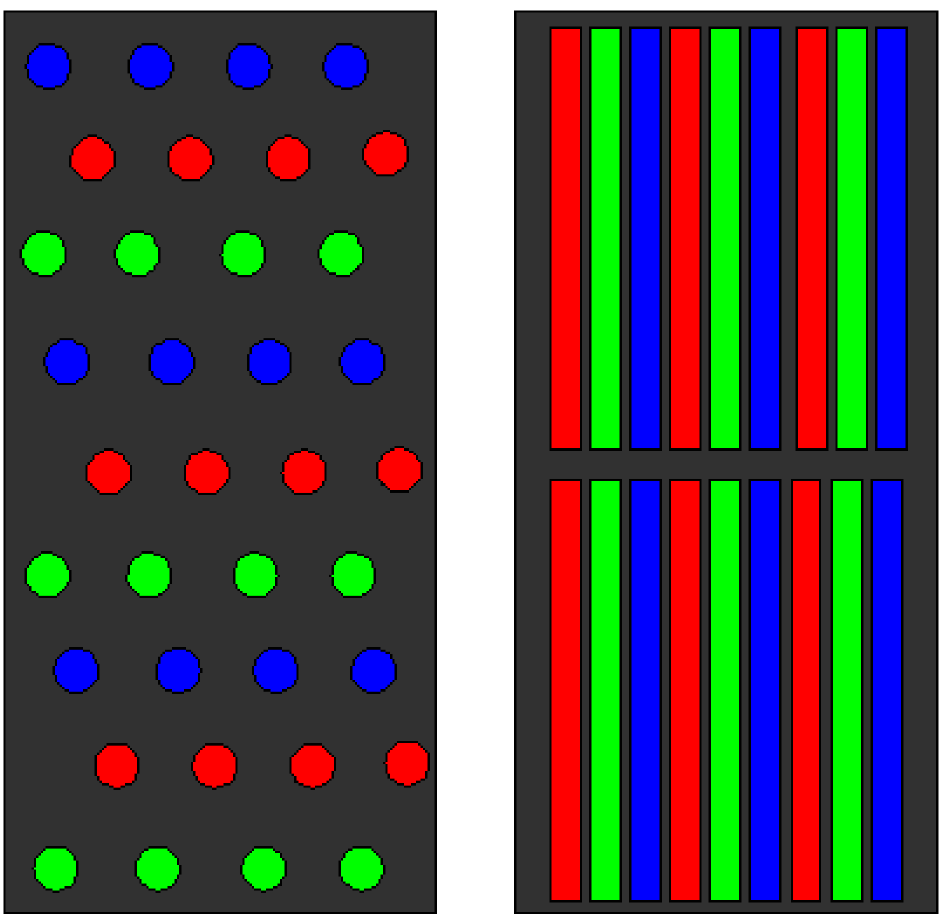
Let us take a closer look at the screens. If you read ads for monitors, you might see many hard to understand technical terms. They may mention many frequencies and dots and pitch?

Note: In many ads, these terms (frequencies, etc.) can appear mixed and unclear. Therefore, be critical when you read monitor data.

Trinitron or Invar

When we talk about traditional screens, there are two primary types of tubes. The best use the so called Trinitron tube. That is a technological principle, which was patented by the Sony company. Since the patent has expired, there are now some clones (ChromaClear, SonicTron etc.).

In the Trinitron screens, the light sensitive pixels on the inside of the tube are placed in a vertical grid, while traditional screens have round masks for the color dots. With the grid mask, you can achieve denser coverage and thus more color saturated images. Here is an attempt to illustrate the difference between those masks:



Traditional mask   Trinitron

The Trinitron screens are generally very high quality. Since the Trinitron tube is more expensive than the traditional Invar tubes, manufacturers also include better control electronics in the Trinitron tubes. That increases their price somewhat, but that money is well spent!

The only disadvantage of the Trinitron (besides price) is the thin lines, which run across the screen. They are visible wires, which contain a grid. In daily work, you will not notice them, but rather enjoy the pleasure of an extremely fine and sharp image.

Invar for contrast

The traditional screen can provide more contrast than the Trinitron screens, which is important in some technical applications. But for ordinary use – in home and offices, where you would typically choose 15" or 17" screens – the Trinitron screens are an obvious choice. Of course, they cost a little more than traditional types, but there is a marked difference in the visible quality. You will experience a much better screen image with a Trinitron tube, no doubt about that!

The horizontal scan frequency

The most important factors are maximum resolution and refresh rate. The screen must be able to deliver an image in a suitable resolution (depending on screen size) and at a good refresh rate (75 HZ or more). The screen can display many different image types – in various resolutions and refresh rates. The interesting point is the *maximum* refresh rate at different resolutions.

These data are often reported together in a number, called the *horizontal scan frequency.* The number is measured in KHZ and it is very important. Basically, the horizontal scan frequency is calculated from resolution and refresh rate. As an example, an 800 x 600 resolution at 75 HZ gives a horizontal scan frequency of 60 KHZ. You cannot calculate the number yourself. Also it varies slightly from screen to screen.

Here are examples of horizontal scan frequency. As I said, the numbers can vary slightly from screen to screen, but they are still in the same ball park:

Different performances		
Resolution	Refresh rate	Horizontal scan frequency
640 x 480	60 HZ	31.5 KHZ
640 x 480	72 HZ	37.8 KHZ
800 x 600	75 HZ	46.9 KHZ
800 x 600	85 HZ	53.7 KHZ
1024 x 768	75 HZ	60.0 KHZ

1024 x 768	85 HZ	68.8 KHZ
1152 x 864	85 HZ	77.6 KHZ
1280 x1024	75 HZ	80.0 KHZ
1280 x 1024	85 HZ	91.2 KHZ

The best performance is provided at the highest refresh rate. The amount of resolution depends on screen size and user habits. In all cases, it would be foolish to run the screen at 31.5 KHZ. That will not at all utilize its capacity.

The multisync screen with digital control

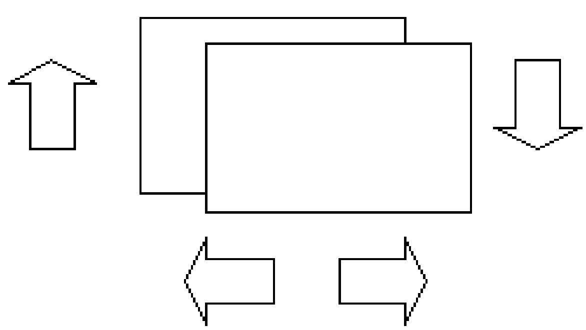
All modern screens are of the *multisync* type. This means, that the screen adjusts itself to the signals received. The individual model has a minimum and maximum horizontal scan frequency. As long as the signals are received within that spectrum, it adjusts itself to the signals.

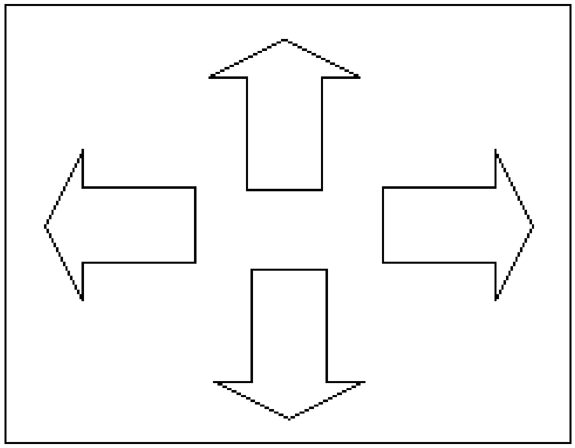
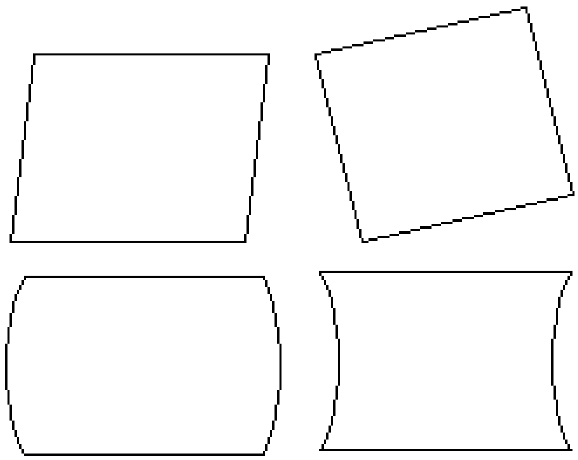
When the screen receives signals at any given frequency, these signals must be adjusted to fill the screen 100%. That is done through the *digital controller* found in modern screens. Older screens would show a clear black border surrounding the image, whenever the resolution was changed to, lets say 800 x 600 and that is very irritating.

To enable adjustment to maximum screen utilization, the screen must have digital controls electronics. These adjustments are made on the screen control panel. We are talking about:

- Horizontal and vertical size, to have the image fill the maximum usable screen area.
- Horizontal and vertical positioning, to center the image.
- Compensation for trapezoid and pin cushioning.
- Colors and light intensity.

The adjustments can look like this:

Adjustments	Symbols
Horizontal and vertical position.	

Horizontal and vertical size.	
trapezoid and pin cushioning.	

Often screens are preset to a choice of different possible adjustments. In these preset conditions, the image will immediately appear perfect.

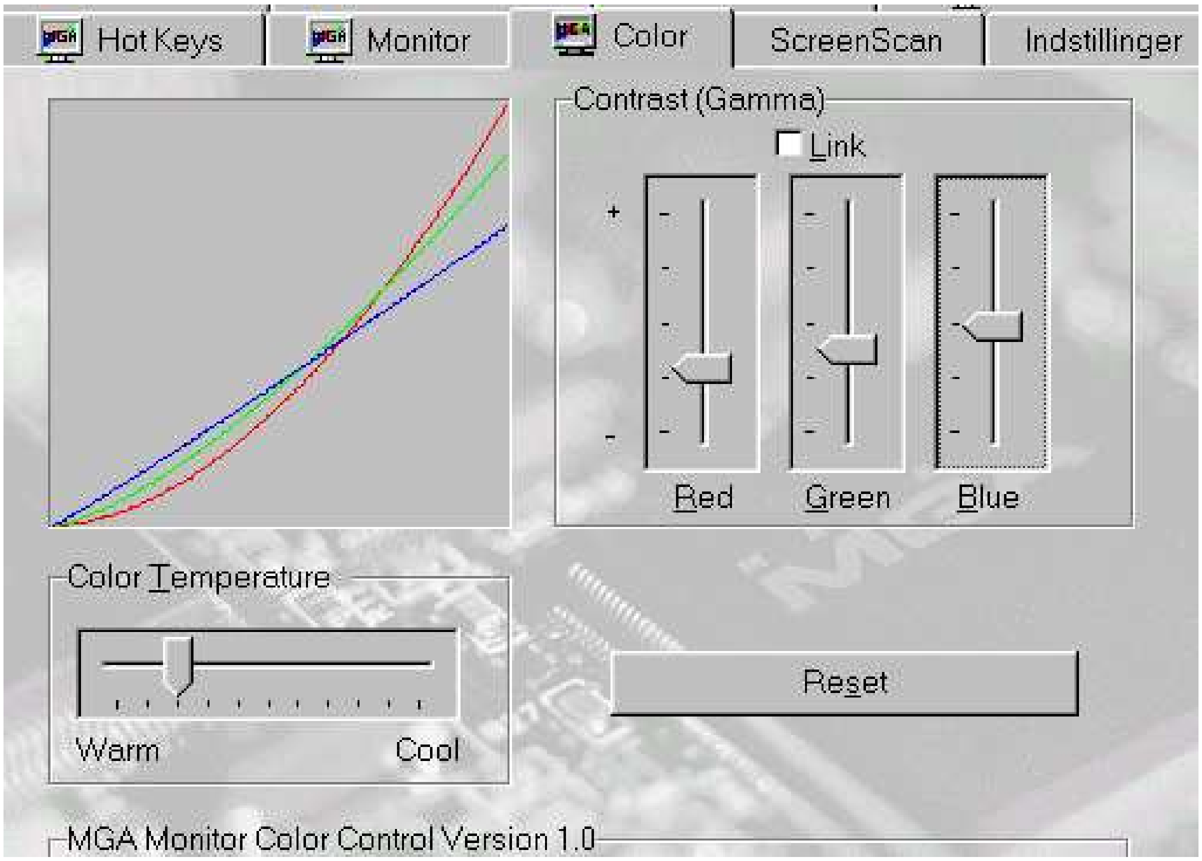
However, when you set up a monitor to work under non preset conditions you have to adjust the image yourself. Once that is done the monitor will remember your settings.

There are no international standards for the design of these digital controllers. They are quite different from monitor to monitor and not all easy to work with. However, working with adjustments is a minor problem, relative to other monitor qualities.

Color adjustments

The screen can show the colors in different heat ranges. The better screens with digital controllers usually have at least two temperature ranges to choose from. I prefer 6500 degrees. 9300 is somewhat colder.

Similarly, some video cards can adjust the screen color temperature like here Matrox:



You should try the different color temperatures. They have a significant effect on the image appearance.

Aperture grill pitch

Often you see the term dot pitch or *aperture grill* pitch. It is measured in millimeters. The number indicate the average distance between individual screen dots. The smaller the better. That provides a finer grain screen. For large screens (21"), the dot pitch can be 0.31 mm or 0.28 mm.

Otherwise, a dot pitch of 0.28 mm or 0.25 mm is considered sufficiently good for ordinary 15" and 17" screens. A few monitors offer 0.22 mm dot pitch.

Screen savers

Early monitors had low quality phosphor coatings. That could result in a screen image to "burn-in" if left unattended. You could clearly see that in work places, where the PC was used for only one program. That program image remained clearly on the screen, after the PC was shut down.

That led to screen savers. In my recollection, Norton's Commander was one of the first of this kind. After a selected number of minutes without activity, the screen switches to moving stars, as if you were flying through space. This prevents the regular image from burning in.

Screens have improved a lot since then - the screen image will not "burn in" in a modern screen. At he same time, screen savers have developed into an art form of their own. Windows 95 is born with a number of choices in screen savers. Also, many programs include a screen saver or two as an extra *feature*. Some provide a series of images, such as "celebrity cars," showing movie celebrities with their fancy cars.

Use the screen savers. They can spice up day-to-day work.

Environmental standards

Screen radiation is a pollutant. There is no concrete evidence that screen radiation can cause illness. However, artificially generated radiation must be unwelcome in our environment. Consequently, industry standards have been developed for acceptable radiation levels.

Since the early nineties, the Swedish MPR-2 standard established limits for monitor electrostatic radiation.

Since came the stricter TCO-92. It limits the permitted amount of low level radiation and establishes standards for electrical and fire safety. TCO means *Total Cost of Ownership*.

Finally, we have TCO-95, which is the strictest standard. Similarly to TCO-92, it also includes regulations on ergonomics (including refresh rates), maximum energy consumption, environmentally friendly production and recycling facilities. The best screens comply with this standard. Screens adhering to the TCO standards are more expensive. Obviously since they are better screens.

The flat TFT screens do not emit *any* radiation at all and they consume considerably less energy than the radiating screens. This is another indication that TFT may be the standard screen of the future.

The VESA DPMS system is an energy saving technology, which includes both screen and video card. A modern 17” screen consumes about 100 watt in normal use. With DPMS the screen switch to two energy saving modes. First, power consumption drops to 25 watt and finally again drops to 8 watt.

LCD screens

The big, heavy traditional monitors will eventually be phased out. They will be replaced by the thin and "soft" LCD (Liquid Crystal Display) screens! It may be a few years before this technology will be dominating, but it is bound to happen. The LCD screens are excellent, and actually they are already available.

The LCD screen is flat, since it contains no cathode ray tube (CRT). In stead the screen image is generated on a flat plastic disk. Here you see a Siemens Nixdorf 3501T. It produces a sharp high resolution image - better than any others I have seen:

LCD screens are also called "soft" screens, since their images have a softer quality than those from traditional CRT monitors. The image does not flicker thus causing less eye strain. People who have become accustomed to these soft images will not return to the traditional monitors. At the same time the LCD screen is by far the most environmentally safe product. These flat screens emit *zero* radiation, and they consume significantly less power than the traditional monitors. Another reason to expect LCD screens to become the monitors of the future.

No refresh-rate

A big advantage in the LCD screen is that it does not flicker. Traditional CRT monitors flicker all the time which is not ideal. Of course the best CRT monitors have a nigh refresh rate (85 Hz or more), which provides a very stable image with no noticeable flicker. But the LCD screen does not flicker at all. They have a refresh rate of 0 Hz!

Modern research has shown that a steadily illuminated screen image is a very important element in a good work environment. The eye responds to all light impressions, and the brain interprets all light impressions continually. When a mediocre monitor flickers, the brain will continually receive superfluous light impressions "noise" to sort out. Thus the brain works permanent overtime interpreting the screen flicker. No wonder that people get tired from watching their monitors.

Digital graphics cards

A special graphics card comes with the 3501 screen. It can be used both with traditional monitors and LCD screens, and you can actually connect to monitors at a time! You save by getting a quality graphics card included in the cost of the LCD screen. The big advantage is that the adapter for the LCD screen is straight digital. The conventional graphics adapters convert the digital signals to analogue signals, which causes some loss of quality. The digital adapter eliminates these distortions completely, which results in a very impressive screen image!

Link to Siemens Nixdorf about LCD screens!

Luminous plastic

The monitor technology is advancing very rapidly. The latest development comes from a British invention LEP - *Light Emitting Plastic*. It is an ordinary thin, flexible plastic (polymer), which is sandwiched together with a thin film of indium-tin oxide and aluminum. Thin-film transistors control the oxide layer, causing the huge plastic polymer molecules to become light emitting.

These LEP screens should have these advantages:

- They are completely flat and lightweight.
- They consume only small amounts of electric power.
- They do not require background illumination, which the LCD crystals do.
- They emit light, which is visible from all angles of view.

These screens are not expected to be available before year 2002, but there are clear indications that they will come. Currently work is being done with prototypes, which have a resolution of 200 dpi. That corresponds to a resolution of 2200 X 1600 pixels in a 15" screen. So we can look forward to an extremely high screen resolution.

I would like to fantasize about future Coca-Cola bottles with a built-in video display in the plastic bottle! By the way, these polymer-plastic materials are finding their way into other parts of the data processing technology. Work is being done on developing different storage media, hard disks in terabytes size and RAM modules based on polymers. These "organic" storage media should also be significantly cheaper to produce that the traditional products.

See Cambridge Display Technology

Read about video cards in Module 7b.

Last revision: 20. May 1998. To Overview.

Click & Learn. Module 7b. WWW.MKDATA.DK

The video card

The video card is just as important as the screen – and more often overlooked.

Two components

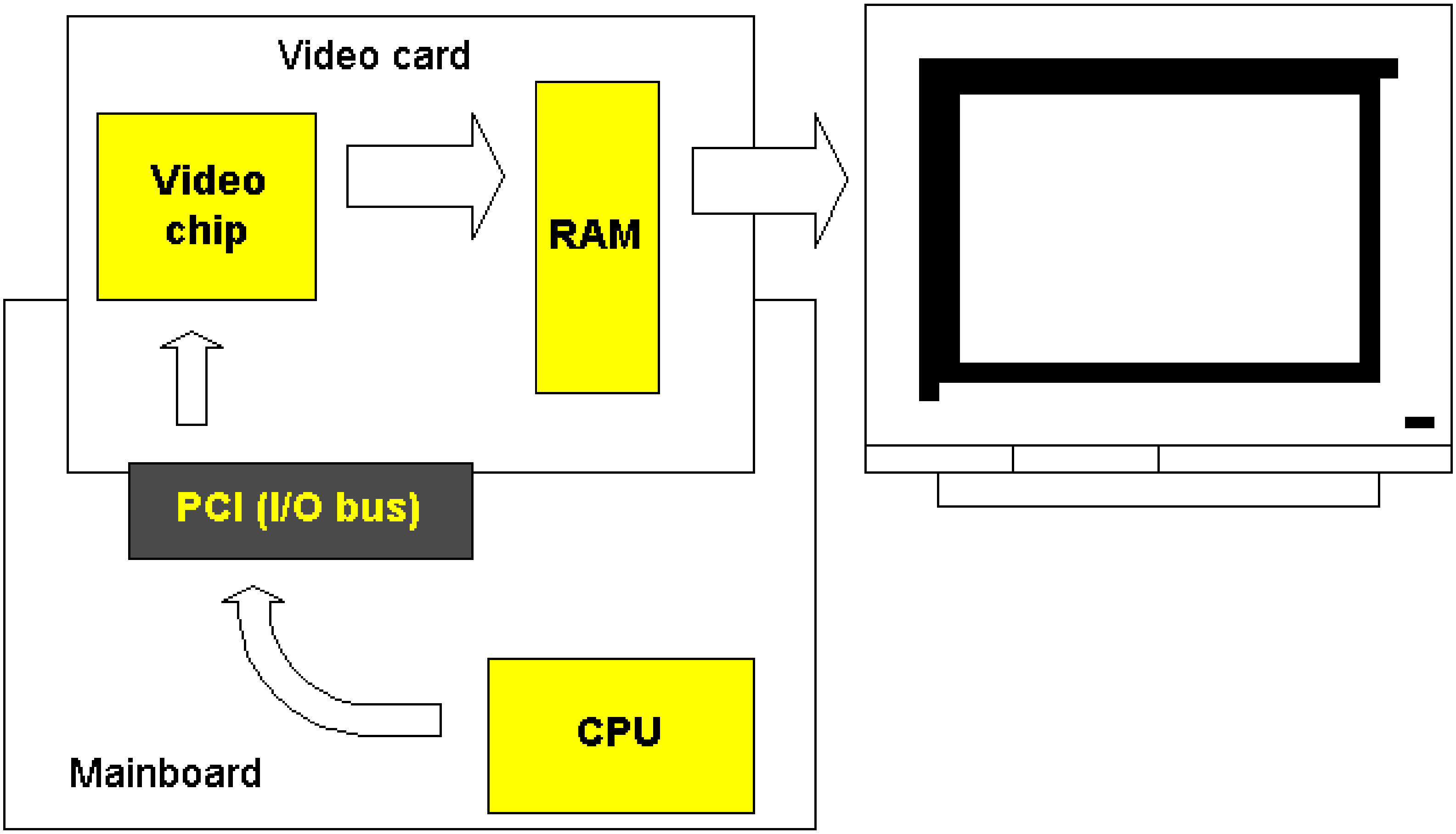
A video card is typically an adapter, a removable expansion card in the PC. Thus, it can be replaced! The video card can also be an integral part of the system board This is the case in certain brands of PC's and is always the case in lap tops. I have a clear preference for a replaceable video card in my stationary PC.

Regardless of whether it is replaceable or integrated, it consists of three components:

- A *video chip* of some brand (ATI, Matrox, S3, Cirrus Logic, or Tseng, to name some of the better known). The video chip creates the signals, which the screen must receive to form an image.
- Some kind of RAM (EDO, SGRAM, or VRAM, which are all variations of the regular RAM). Memory is necessary, since the video card must be able to remember a complete screen image at any time.
- A RAMDAC - a chip converting digital/analog signals.

The video card supports the CPU

The video card provides a support function for the CPU. It is a processor like the CPU. However it is especially designed to control screen images.



Heavy data transport

The original VGA cards were said to be "flat." They were unintelligent. They received signals and data from the CPU and forwarded them to the screen, nothing else. The CPU had to make all necessary calculations to create the screen image. Each screen image was a large bit map. Thus, the CPU had to move a lot of data from RAM to the video card for each new screen image.

The graphic interfaces, like Windows, gained popularity in the early nineties. That marked the end of the "flat" VGA cards. The PC became incredibly slow, when the CPU had to use all its energy to produce screen images. You can try to calculate the required amount of data.

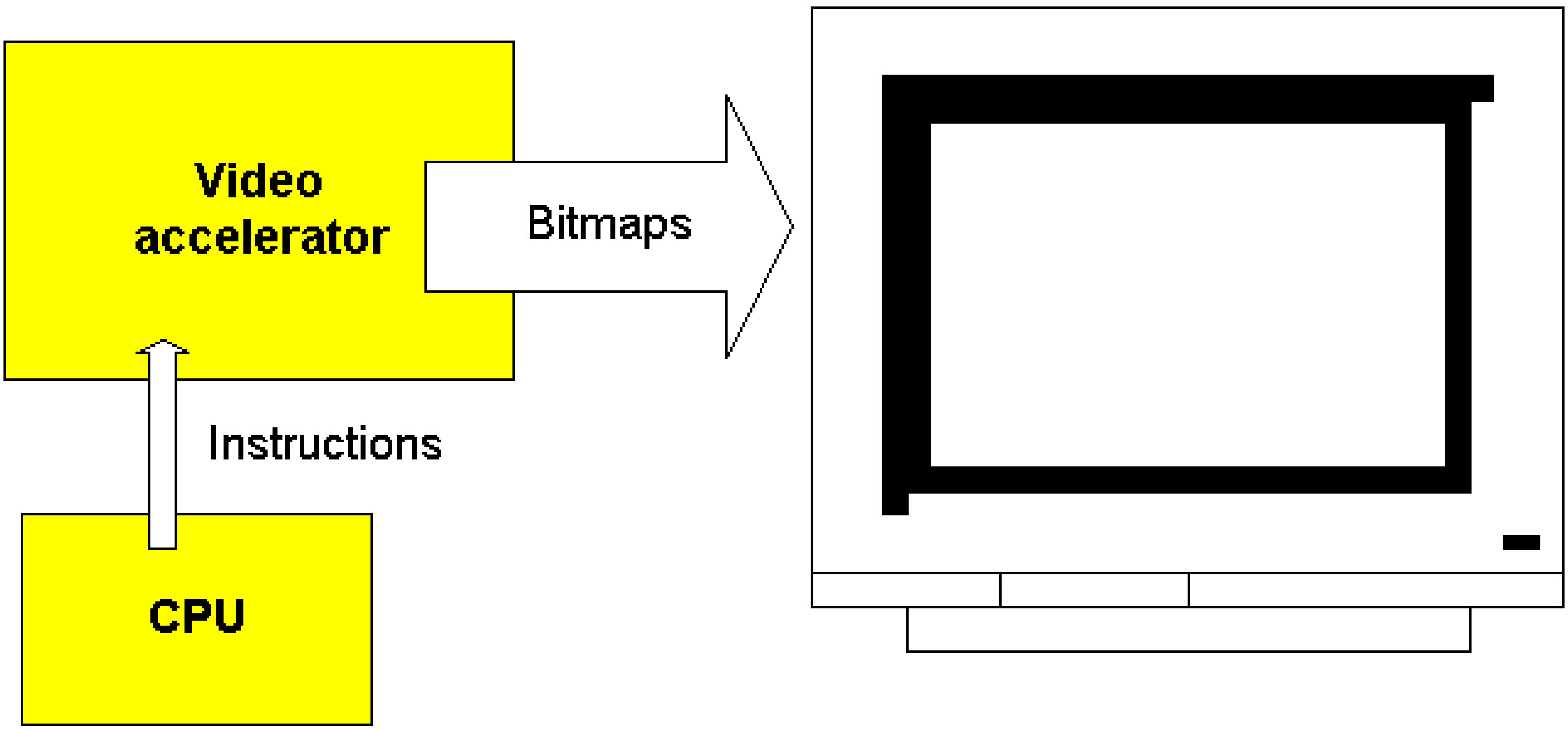
A screen image in 1024 x 768 in 16 bit color is a 1.5 MB bit map. That is calculated as 1024 x 768 x 2 bytes. Each image change requires the movement of 1.5 MB data. That zaps the PC energy, especially when we talk about games with continual image changes.

Furthermore, screen data have to be moved across the I/O bus. In the early nineties, we did not have the PCI bus, which can move large volumes of data. The transfer went through the ISA bus, which has a very limited width (read in module 2b about the busses). Additionally the CPU's were 386's and early 486's, which also had limited power.

About accelerator cards

The solution to this problem was the *accelerator* video card, which appeared in the early nineties. All newer cards are accelerated and today they are connected to the CPU through high speed busses like PCI and AGP.

With accelerated video cards, Windows (and with that the CPU) need *not* calculate and design the entire bit map from image to image. The video card is programmed to draw lines, windows, and other image elements. The CPU can, in a brief code, transmit which image elements have changed since the last transmission. This saves the CPU a lot of work in creating screen images. The video card carries the heavy load:



All video cards are connected to the PCI bus, this way providing maximum data transmission. The AGP bus is an expanded and improved version of the PCI bus - used for video cards only. AGP will be the new standard in coming years.

RAM on the video card

Video cards always have a certain amount of RAM . Two RAM features are significant:

- How much RAM? That is significant for color depth at the highest resolutions.
- Which type RAM? This is significant for card *speed*.

Video card RAM is necessary to keep the entire screen image in memory. The CPU sends its data to the video card. The video processor forms a picture of the screen image and stores it in the video card RAM. This picture is a large bit map. It is used to continually update the screen image.

Amount of RAM

Video cards are typically available with 1, 2, 4 or more MB RAM. How much is necessary? That depends primarily on how fine a resolution you want on your screen. For ordinary use, 16 bit colors are "good enough." Let us look at RAM needs for different resolutions:

Resolution	Bit map size with 16 bit colors	Necessary RAM on the video card
640 x 480	614,400 bytes	1 MB
800 x 600	960,000 bytes	1,5 MB
1024 x 768	1,572,864 bytes	2 MB
1152 x 864	1,990,656 bytes	2,5 MB
1280 x 1024	2,621,440 bytes	3 MB
1600 x 1200	3,840,000 bytes	4 MB

Note that the video RAM is not utilized 100% for the bit map. Therefore, 1 MB *is not* enough to show a 800 x 600 picture with 16 bit colors, as the above calculation could lead you to believe.

Most video cards come with 2 or 4 MB RAM. The table clearly indicates that 1 MB RAM has very limited use. If you stay with 1024 x 768 or less, and with 16 bit colors, 2 MB RAM is quite sufficient.

If you have a large screen – 17" or more, you should seriously consider getting 4 MB RAM on your video card. I am very enthused about the little-known 1152 x 864 resolution. It is very pleasant to work with, but you need 4 MB RAM on the video card.

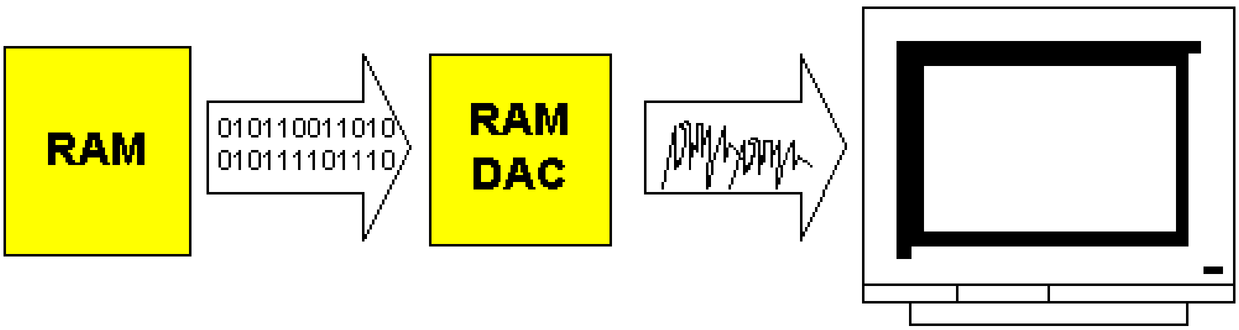
VRAM

Briefly, in principle all common RAM types can be used on the video card. Most cards use very fast editions of ordinary RAM (like EDO).

The high end cards (like Matrox Millennium II) use VRAM (*Video* RAM). This is a RAM type, which is only used on video cards. In principle, a VRAM cell is made up of two ordinary RAM cells, which are "glued" together. Therefore. you use twice as much RAM than otherwise. VRAM also costs twice as much. The smart feature is, that the double cell allows the video processor to *simultaneously* read old and write new data on the same RAM address. Thus, VRAM has two gates which can be active at the same time. Therefore, it works significantly faster.

RAMDAC

All graphics card has a RAMDAC chip converting the signals from digital to analogous form. Traditionally the monitors works on analogous signals. The PC works with digitalized data which are sent to the graphics adapter. Before these signals are sent to the monitor they have to be converted into analogous output and this is processed in the RAMDAC:



A good RAMDAC:

- External chip, not integrated in the VGA chip
- Clock-speed: 200 - 250 MHZ.

Flat LCD-monitors works on a digital input. My own Siemens Nixdorf 3501T came with a *digital graphics adapter*, where the RAMDAC isn't used. There is a RAMDAC on the card since it is capable of controlling a traditional CRT-monitor simultaneously with the LCD-screen.

Cards and chips

There are many manufacturers of video cards and accelerator chips. Some produce both cards and chips, while others only make one or the other. I have tried a lot of different cards. The tables below illustrate my personal evaluation. Some may disagree with my evaluation.

First the best known video cards:

Make	**Quality and price**
ATI	Medium - High

Cirrus Logic	Low
Matrox	High
S3	Medium
Tseng	Low - Medium

Here we see video card manufacturers:

Make	**Quality and price**
ATI	Medium - High
Diamond	Medium (High)
Matrox	High
Number Nine	High
Orchid	Medium (High)
STB	Medium
Britek/Viewtop	Low

You can use these tables, when you buy a PC and/or video card. Make sure to start with a quality video card!

You should get a demonstration of the card and monitor you want to buy. Especially if it is in the medium group, I will strongly recommend that you see it connected. Then evaluate the screen image. How sharp is it? Does it flicker? Ask about resolution, color depth and refresh rate. If the dealer cannot answer these questions, I would not trust him. Finally, find out which *driver* the card needs. Read on...

The driver – almost the most important part

The difference between good and mediocre cards is clearly visible in their software. The companies ATI, Matrox and Number Nine deliver excellent drivers with their cards. They allow their cards to provide optimum screen performance.

In contrast I can mention the ET 6000 accelerator chip, which was introduced about a year ago. It has very fine specifications and scored very high in various tests. I bought a couple of cards with that chip, but I could never get them to work properly. The driver programs are poorly written, for example the refresh rate is not adjustable. Such cards are all right for low quality monitors, but not for monitors with high specifications. For these the refresh rate should be adjustable.

Here you see a section of the Matrox video card control box. The driver knows precisely which refresh rates the monitor will tolerate at different resolutions:

Resolution	Maximum Vertical Frequency
640x480	75 Hz
800x600	85 Hz
1024x768	85 Hz
1152x864	85 Hz
1280x1024	85 Hz
1600x1200	70 Hz

Another problem area is in the screen fonts, which come with the driver programs. Screen fonts are models for the letters seen on the screen. There are significant quality variations in this area. Again, ATI, Number Nine, and Matrox are worth mentioning.

Video card is one of the items hardest to choose

It is difficult to choose a video card, because there is such a multitude of different ones. And you can read *test reports* forever. Yet, they may not be particularly useful.

One of the best cards I ever laid my hands on was IBM's XGA-graphics, which unfortunately was never available separately. Once they provided a working Windows driver, this card was in a class by itself.

One of the problems with video cards is that the same graphics chip may be used in both good and inferior video cards. This is especially true for S3 chips, which are used in the best cards from Number Nine and also used in ViewTops discount cards. When a fundamentally good chip is found both in great and in mediocre cards, you cannot select the card based on what chip it uses!

Another problem is in the test methods, which the computer magazines use. They measure exclusively *speed*. Speeds are measured with special programs, which read how fast the screen image can be built, etc. That's fine – but is does not say much about image quality, as perceived by the eye. Is it sharp, bright, not flickering, comfortable? Those are more subjective and abstract qualities, which can never be evaluated by a test program.

You should choose a card based on its specifications. For example, can it deliver a 1024 x 768 image at 85 HZ? It should be able to do that, but not just in theory. It must also be able to do that in real life. Here is where the *driver* comes in.

Video cards and chips, different brands

Let me comment on a couple of brands:

The Canadian firm ATI was among the first to produce accelerated video cards, when the graphic milieus came on board in the early nineties with Windows. The company's first chip was called *mach 8*. That had an 8 bit graphic processor, which was extremely fast relative to others at the time. But they were extremely expensive!

Later, ATI presented *mach 32* and *mach 64*, which were 32 bit and 64 bit graphic processors, respectively. All the way, ATI has produced solid video cards with good quality drivers. Today they are available in many price ranges, including low cost editions. You will never go wrong with an ATI card.

The Matrox company is also Canadian, and originated from ATI. They also make excellent cards with their own accelerator chips. They only make a few models. Regardless of which Matrox card you buy, it is an excellent product. Matrox comes with good drivers. Obviously to be recommended.

Tseng has made graphic chips for many years. In the good old DOS days, an ET 4000 card was one of the best on the market. It was equipped with Tseng's ET 4000 chip, which was excellent for DOS usage. Since came the somewhat overlooked ET4000/W32 chip. I had good experiences with that on some low cost ViewTop cards. Tseng's latest chip is ET 6000, which is mostly sold as a discount card. Not recommended.

S3 is a big name in graphic chips. They do not manufacture their own cards, but their chips are used in numerous cards. companies like IBM, Diamond, Number Nine, and ViewTop/Britek use S3's different accelerator chips with widely varying results. A small S3 Trio 64 chip is mounted in IBM's PC 300, which on paper is not very powerful. Yet, it produces an incredible fine image. Thus, the quality depends just as much on other video card design features as on the accelerator chip. Can you then recommend S3 based cards at all? Yes, if they are made by a respectable manufacturer, who includes a quality driver. And that has to be tried out in practice.

3D graphics

3D images, where you can move around in space, is a technology, which is expanding to the PC world. Ordinary PC's to day are so powerful, that they can actually work with 3D environments.

Ordinarily, our screen images (such as in Windows 95) are two dimensional. But we know 3D effects from movies and from some computer games.

Over the last few years, more and more 3D standards have arrived in the PC market. That includes:

- *VRML*, which provides 3D space on the Internet.
- Direct 3D API, which enables programming of 3D games for Windows 95

The 3D technologies are of no consequence for ordinary office programs. They are used in certain games, like Quake and others. Actually, the ordinary video cards are optimized to show 2D images.

2D cards can construct 3D movements, but it will take time to bring the images to the screen. That is because of the very complex calculations needed. Therefore, hardware accelerators have been designed. They can give drastic improvements. Also, special functions are included in the video card chip, allowing it to calculate 3D movements lightning fast.

The most convincing 3D performance are achieved with a special 3D card like Orchid Righteous 3D. It is a specialty card, used only for 3D presentations. You will need a regular video card also. You simply draw a cable from the 2D video card to the 3D card, where the screen is also connected. But many ordinary video cards, like Matrox Mystique 220, also have quite effective 3D accelerators built in.

Intel has gone into the 3D business with their successful i740 chip. Later "Portola" is expected for the katmai processors. It should improve 5 times to i740.

Spend money on the screen image

When you have read all these technical explanations, the *choosing* and *buying* remains. In my mind, the screen image is without doubt extremely important for your daily work. It is also an area with vast quality variations. Often PC's are advertised with a very cheap monitor and the very cheapest S3 video card. Many would be happy with this equipment. The PC works fine and they may never have seen a high quality screen image.

I will strongly recommend that you invest a little more, to get a better video system. Specifically I would recommend a 15" or 17" Triniton monitor. It could be a Mag or Nokia. Both make excellent Trinitron screens at reasonable prices. Combine the monitor with an ATI, Number Nine or Matrox video card and you will have a good video system!

The screen image and Windows 95

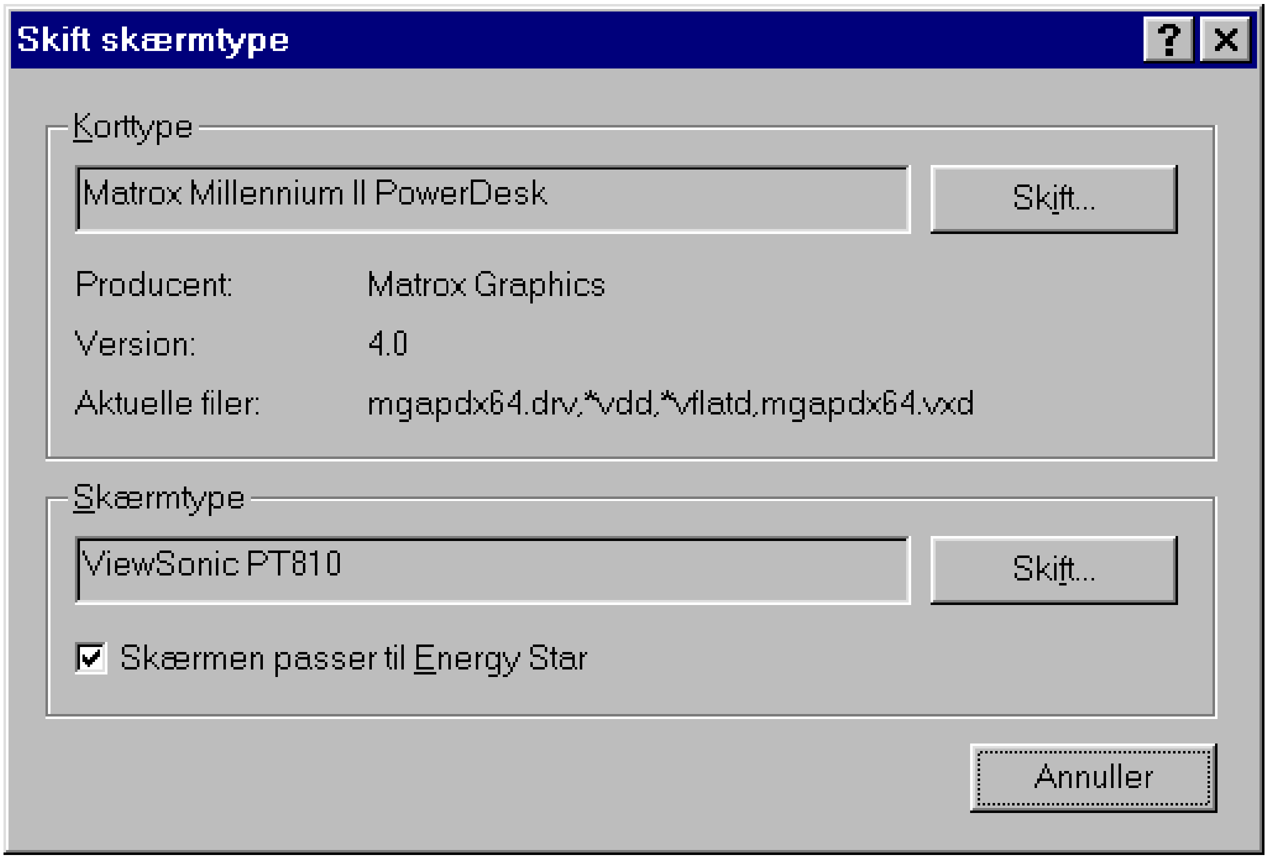
Once you buy *both* a good monitor *and* video card , you have to make them work together. That is done in Windows 95 through driver programs. That part of the installation is extremely important, it requires attention.

If you leave it to Windows 95 to install the necessary drivers, the result may be mediocre. Windows 95 is so smart, so along the road it will find your hardware,. And Windows 95 will install drivers, when it encounters new software. Often some *standard drivers are installed*. They will make the software work, but no more.

If you care about your screen image, you must make sure you have the correct drivers installed. We are talking about two drivers:

- Driver for the video card (the most important).
- Driver for the monitor.

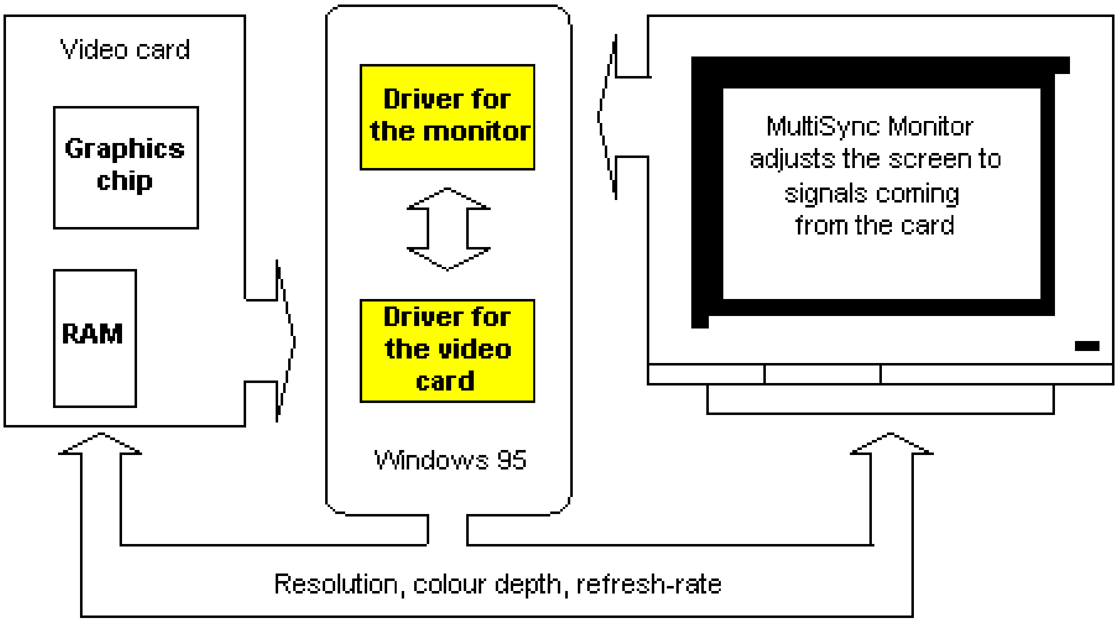
Both can be found in Windows 95, in *my computer -> control panel -> display -> settings -> advanced properties*. Here you see my current settings:



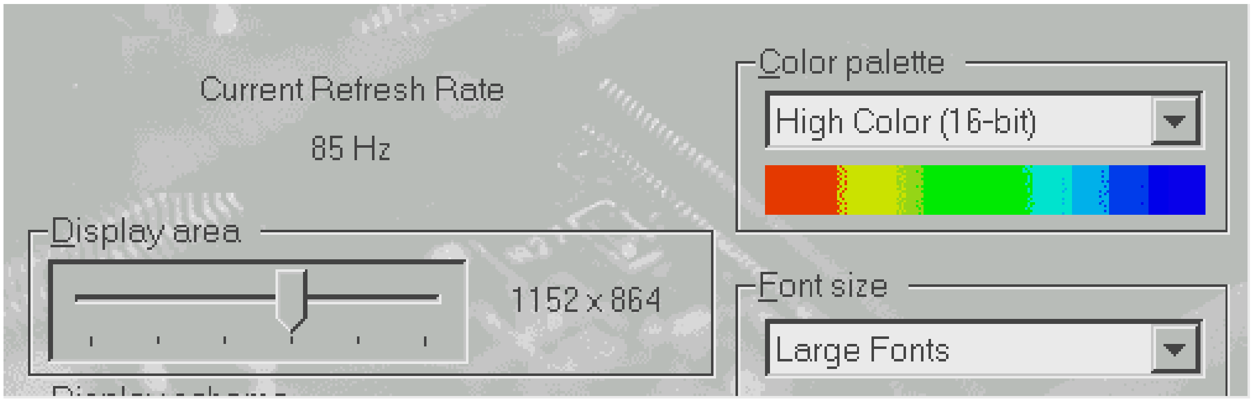
On top you see the card and below the monitor. Both are named, and the optimal drivers are installed. This allows Windows 95 to get a full picture of the video system. Then the video card can deliver the optimal signals to the screen.

I have personally ruined a 17" monitor by changing the video card. I adjusted the new card to deliver precisely the maximum ability of the monitor - according to the specifications. However, the monitor was a few years old. It had always run at a lower resolution and refresh rate, to which it must have adjusted itself. It did not work out - the electronics burned out!

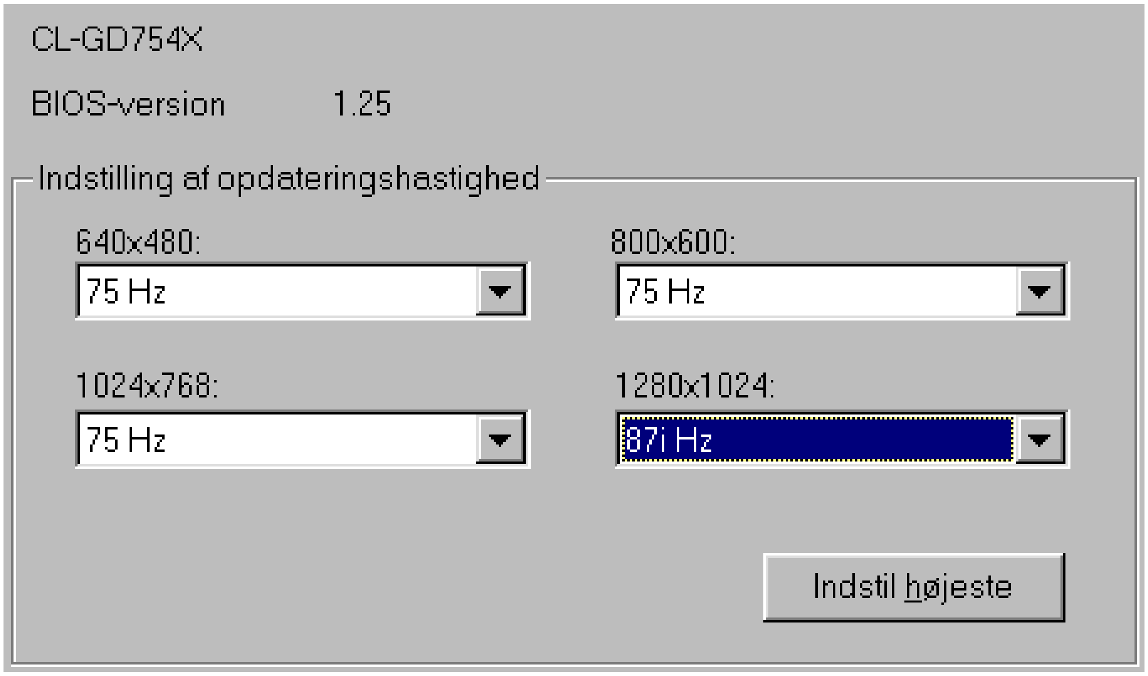
The Windows drivers link video card and monitor together, and make them cooperate with each other.



The standard driver in Windows 95 cannot adjust refresh rates. Therefore, a driver has to be installed to exercise this option. Here is a Matrox Millennium II video card, with its own dialog boxes installed:



And here are the adjustments from my notebook, which has an adjustable Cirrus Logic video chip:



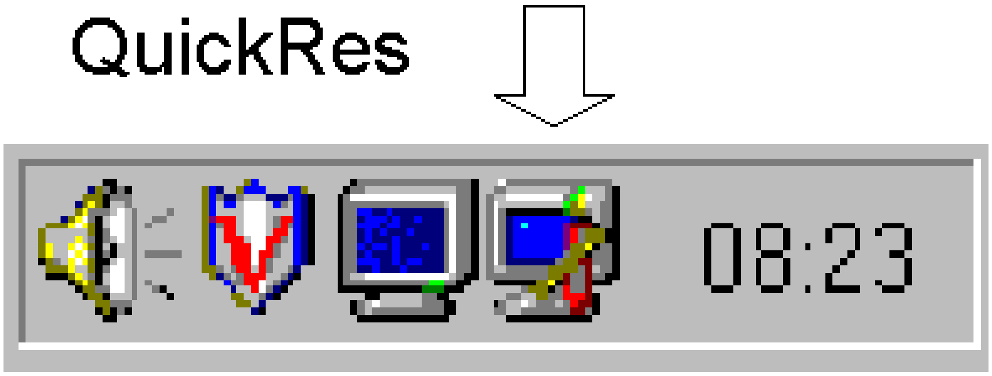
You *need* to install a driver program, which works specifically with your video card. Otherwise, you are guaranteed not to utilize your video card efficiently. Very few dealers seem to understand this concept. Nearly all PC's are sold with Windows 95 standard driver installed and the video system will render absolute minimum performance!

DCC

VESA DDC (Display Data Channel) are technologies, which should allow the video system to find the optimum adjustment, through communications with the video card. I do not think it is quite working yet.

QuickRes

If you want to experiment with different screen resolutions, you can download the program: QuickRes.exe. It is a small Windows 95 utility application, which you then have to run (double click on it). Then, the program will appear as a small icon in the lower right corner of your screen:



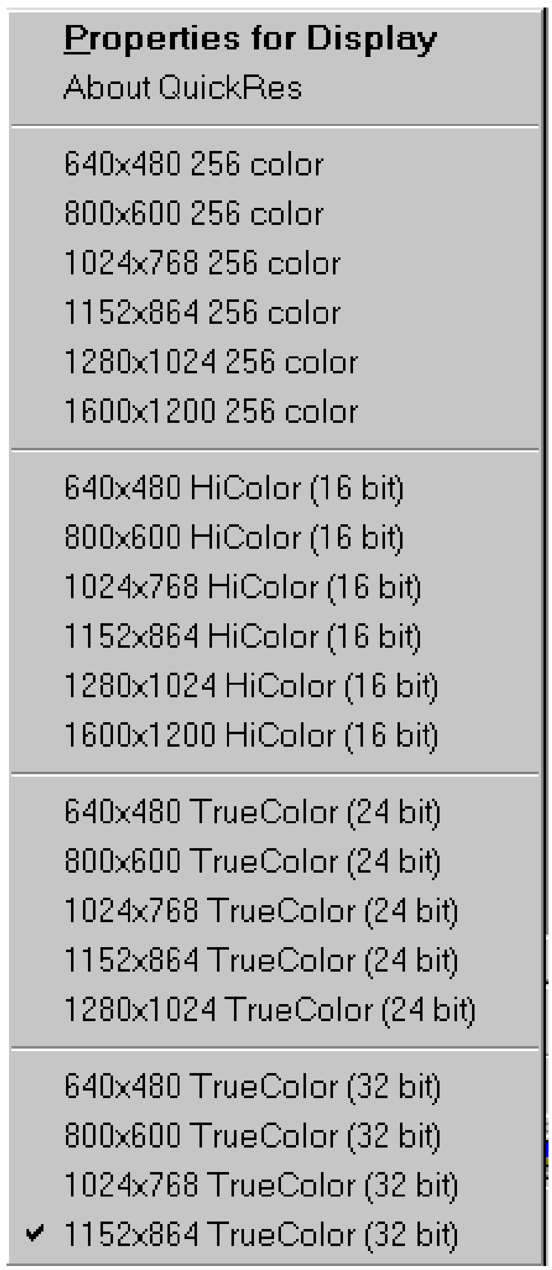
When you right click on the icon, a menu will open showing all the resolutions and color depths you can choose from on your PC:

Here you can see, that the resolutions on my PC go from 640 x 480 up to 1600 x 1200. Color depth goes from 8 bit (256 colors) to 32 bit.

Note that the maximum color depth at 1600 x 1200 resolution is 16 bit and at 1280 x 1024 it is 24 bit.

Only the 1152 x 864 resolution can be seen in full 32 bit colors.

This limitation is because my video card only has 4 MB RAM installed. However, that has no practical consequences for me.



QuickRes is smart, because you can change resolution "on the fly". Normally Windows 95 has to be re started, but here, the screen image just blinks a couple of seconds, then the new resolution is in place.

Last revision: 8 Jun 1998. To Overview.

Click & Learn. Module 7c. WWW.MKDATA.DK

The sound card

I do not claim to be an expert in sound cards. But I will try to describe what little I know about this technology. The sound capabilities of the PC is a very exciting area - not in the least here in 1998, where radical new designs in sound technology appear feasible. On the first pages I will describe the traditional sound card concept (the Sound Blaster compatible sound card). Then I will describe the new technologies which appear on the horizon.

Introduction

Sound cards have a minimum of four tasks. They function as

- Synthesizer
- MIDI interface
- Analog/digital converter (A/D), when sound is recorded from a microphone.
- Digital/analog converter (D/A), when the digital sounds have to be reproduced in a speaker.

The synthesizer

The synthesizer delivers the sound. That is, the sound card generates the sounds. Here we have three systems:

- FM synthesis, Frequency Modulation
- Wave table
- Physical modeling

The cheapest sounds card use the FM technology to generate sounds simulating various instruments. Those are true synthesizers. The sounds are synthetic – it may sound like a piano, but it is not. FM synthesis are and sound like artificial sounds.

Wave tables

Wave table is the best and most expensive sound technology. This means that the sounds on the sound card are recorded from real instruments. You record, for example, from a real piano and make a small *sample* based on the recording. This sample is stored on the sound card.

When the music has to be played, you are actually listening to these samples. When they are of good quality, the sound card can produce very impressive sounds, where the "piano" sounds like a piano. Wave table is used in Sound Blasters AWE card.

Physical modeling

Physical modeling synthesis has arrived as a third sound producing technology. It involves simulating sounds through programming. The process is supposed to be rather cumbersome, but it should yield a number of other advantages. The Sound Blaster Gold card contains 14 instrument sounds, which are created from physical models.

The basic quality of a sound card can be tested by playing a MID file. Then you can easily hear the difference.

There is also a difference in how many notes (polyphony) can be played simultaneously. If you want to compose your own music on your PC, you use the sounds available on your sound card. The greater works you want to write, the more "voices" you will need. The SB AWE64 card has 64 voices, while SB16 only has 20 voices.

Some sound cards can import new sounds. They are simply downloaded to the sound card, which might have 512 KB (Sound Blaster AWE64) or 4 MB RAM (Sound Blaster AWE64 Gold) available for the users own sounds.

MIDI

MIDI (Musical Instrument Digital Interface) is a specification, which was developed in the 1980's to communicate between synthesizers. Since then MIDI has also become the standard, which allows programs to play music through the PC sound card.

MIDI is a computer standard music format. You write compositions - musical events - in the MIDI format. The MIDI files do *not* contain the sounds but a description of how the music is to be played. The sounds are in your sound card. For example a MIDI sequence can describe the hit on a piano key. The MIDI sequence describes:

- The instrument
- The note

- The strength of the key hit
- How long to maintain the note
- Etc.

The only thing which is not covered is the sound of the instrument - that is created in the sound card, and is totally dependent on the sound card quality.

A MIDI recording is thus a recording of music on "note level," without sound. It is played by a module, such as a sound card, which can generate the sounds of the instrument.

MIDI files do not occupy much space as compared with the pure sound (WAVE files). Therefore they are often used in PC's, on Internet etc.

MIDI interface for keyboards

A musical keyboard can be connected to the sound card with a connector. That is called a MIDI interface. You can buy special PC musical keyboards, or you can use one of the keyboards which are available in music stores. It will work as long as the MIDI connectors match.

You connect your DIN connector to the piano keyboard. In the other end of the cable is a DB15 connector to the sound card. Then you can play from the piano keyboard through the sound card. Of course it requires a program which can handle music, but it works.

I have tried it myself. The Sound Blaster AWE64 Gold comes with the program Cubasis. Once I connected an 8 year old cheap piano keyboard (with built-in rhythm box) to the sound card, and everything worked through Cubasis. The keyboard acted as a "Local Synthesizer" in the program settings.

This keyboard is especially designed for the PC.

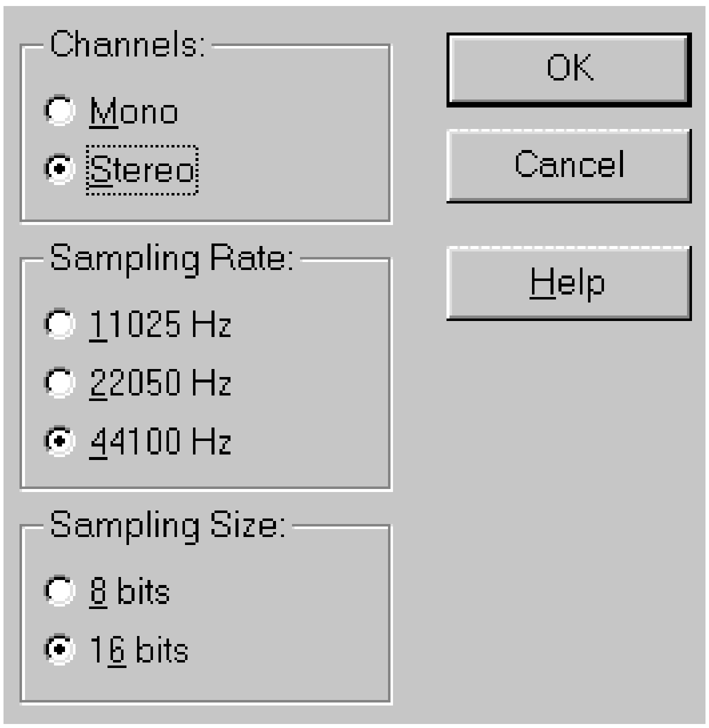
The A/D conversion

When you connect a microphone to the sound card, you can easily record your own voice on the PC. The result is a small WAVE file which holds a digital recording of the sound, which reached the microphone. The sound is analog, the file is digital – the transformation is done in the A/D converter in the sound card.

This sound recording is called a *sampling*. It can be done in various qualities:

- 8 bit or 16 bit sampling
- 11, 22 or 44 KHz (how many thousand times per second the sound will be recorded)
- Stereo or mono

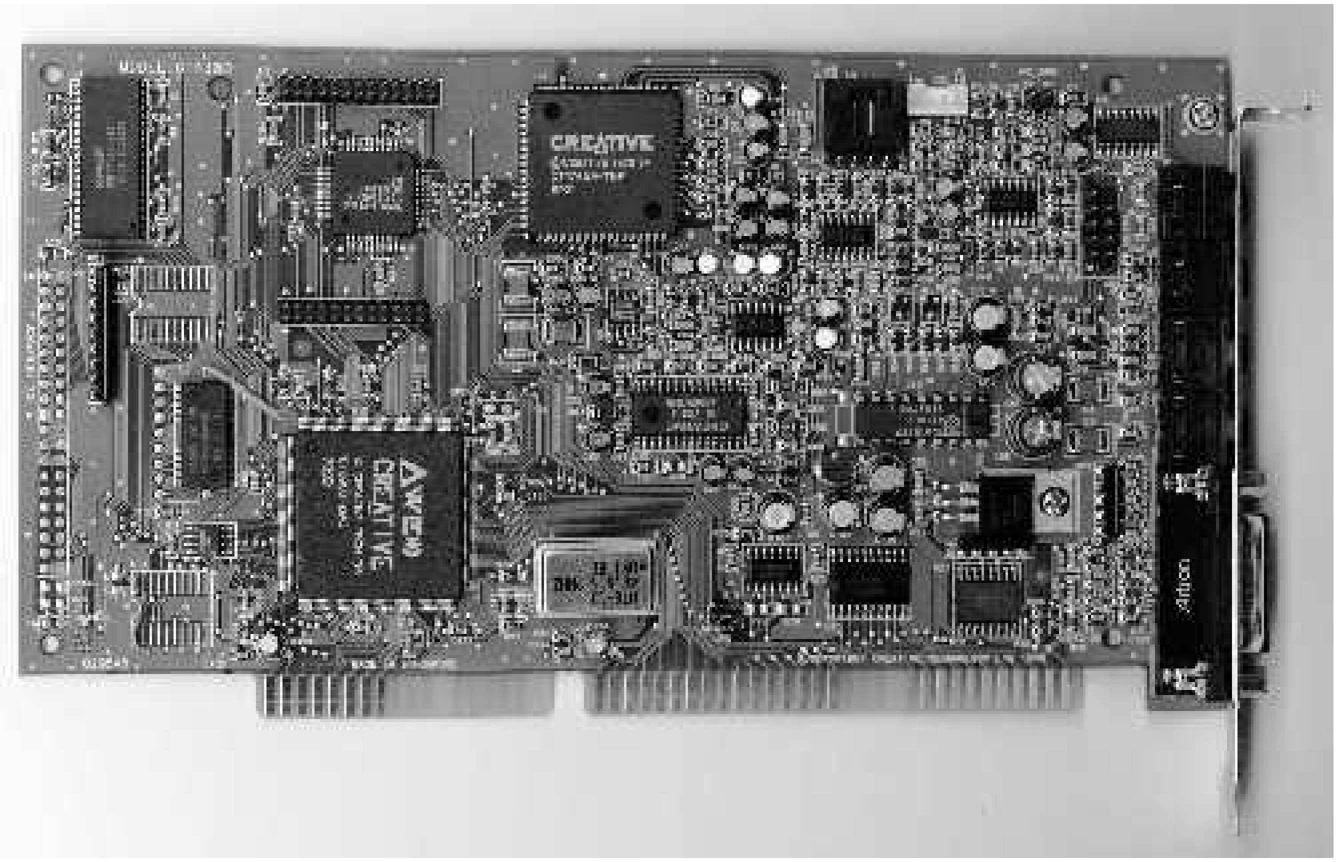
Here you see the settings in a Wave program:



Stereo sampling at 16 bit and 44 KHz gives the best quality, but the Wave files will take up quite a bit more space.

The sound card - an adapter

The sound card itself is an ISA adapter. Here you see an AWE64 Gold card:



The connectors may look different on different sound cards, but as an example: In the back of the AWE64 Gold card you find connectors to:

- Microphone input, a jack
- Line input, a jack
- Two phone jacks for active speakers
- A DB15 jack for MIDI or joystick.

Sound cards typically have a 2 Watt amplifier built-in. It can pull a set of earphones. The exception is the Gold card, where the amplifier is eliminated. It has no practical significance, since you probably want to attach it to a pair of active speakers.

The sound of the future

Until now PC sound has been totally dominated by the Sound Blaster card. All sound cards had to be compatible with Sound Blaster, or it would not sell. Obviously that is due to the numerous game programs, which require a SB compatible sound card.

The new sound cards break away from the Sound Blaster compatibility. This break involves many facets. Below I will describe some of the tendencies in the sound technology.

Sound over the PCI bus

Newer sound cards can be connected on the PCI bus. The SB compatibility requires the old ISA bus, which is really antique today. With PCI you gain different advantages:

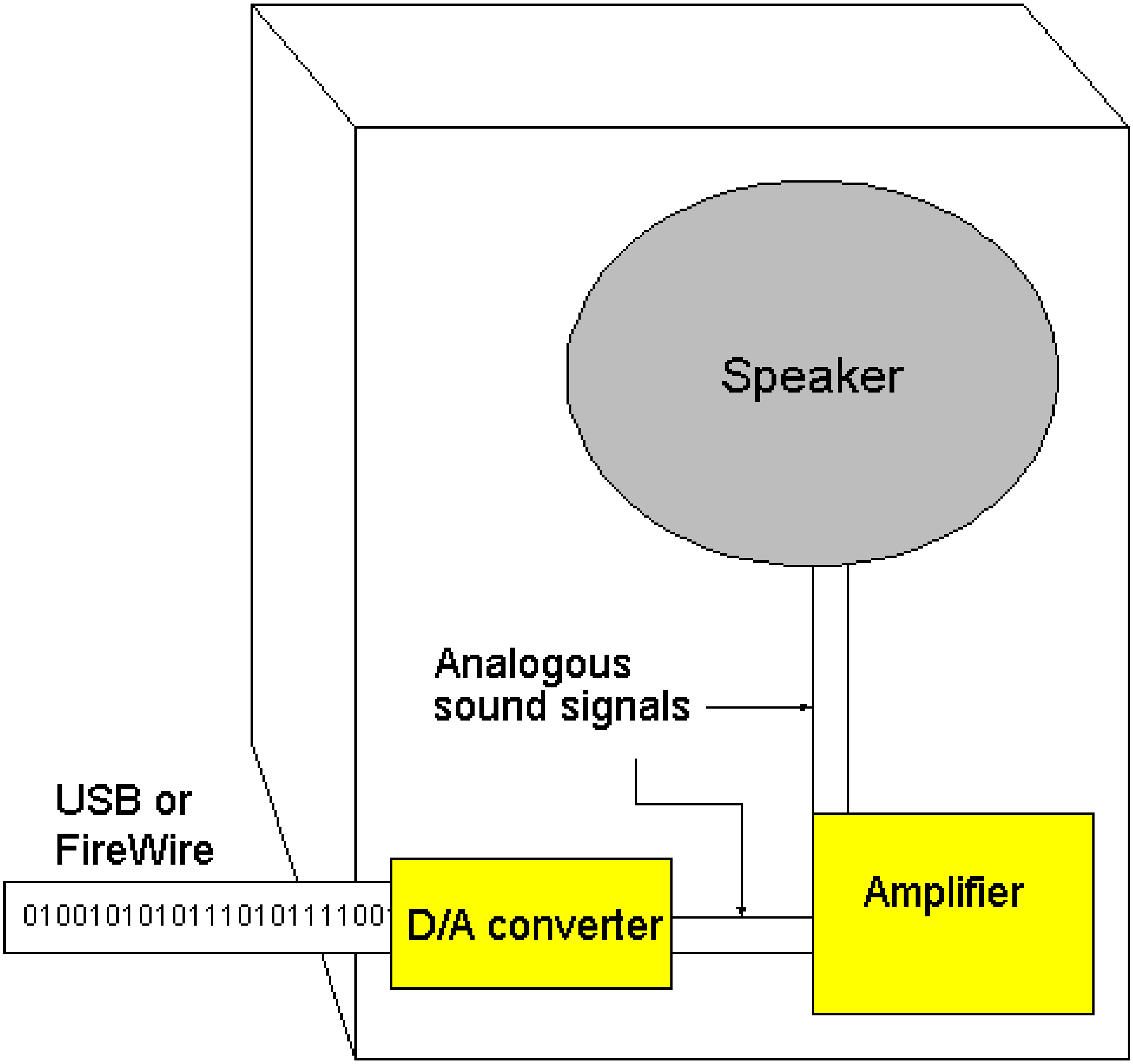
- The IRQ problems disappear.
- Signal/noise ratio can be improved with 5 dB.
- There is sufficient bandwidth (capacity for data transmission).
- The sound card workload for the CPU is less.
- We can drop the ISA bus, which takes up unnecessary space on the PC system board.

The problem in moving the sound to the PCI bus involves existing software. First of all old DOS games, which expect and demand the Sound Blaster card with its IRQ- and DMA number. They will not work with the new cards, unless special solutions are implemented.

Sound over the USB bus

We will experience very high quality sound systems, when the new USB bus is implemented.

The difference is that the sound card in the PC will work totally digital. The sound signals are in digital form when they are sent out on the USB channel:



Inside the PC there is a lot of electric (static) interference from many sources. That can affect the integrity of the signals in the sound module. With USB the noise sensitive digital/analog conversion will take place in the speaker, and this results in a superior quality. Both Philips and Altec Lansing produce USB speakers.

In the future we will see Hi-Fi speakers with built-in amplifier and converter, which can receive pure digital signals (via USB). These speakers will randomly be able to interpret data from Hi-Fi equipment, PC, TV/video and other sources. Surely we will also see sound cards and speakers for the FireWire bus, which is somewhat similar to USB.

DOS or DirectX

When so many games are DOS based, it is primarily because of the sound. Under DOS the programmer can modify and manipulate the sound card to a very high extent. It can be controlled very precisely, sounds can be mixed without interruption, and all kinds of effects can be designed. Here DOS proves very effective - the operating system permits direct control of the hardware. The disadvantage is, that the hardware must be totally standardized. This gave the Sound Blaster card its great success.

In Windows all program instructions to hardware are executed through a programming layer (API). For example the first multimedia-API would not allow mixing of sounds. Therefore the music in a Windows-based game had to be cut off, if there was a need for playing such a thing as the sound of an explosion. This put heavy restraints on programming creativity. Consequently DOS based game applications remained long into the Windows era.

DirectX is a set of multimedia API's (*application program interface*) developed for Windows. It is a collection of programs which enable much improved *low level control* over the hardware in games and other multimedia applications. DirectX has now reached version 5.0 and includes:

- DirectDraw
- DirectSound
- DiectSound3D
- DirectPlay
- DirectInput
- DirectSetup

All of these programs are designed to enable all possible image and sound effects. The advantage of DirectX is that the applications can be written directly to Windows and simultaneously get maximum hardware control. With this we should finally have eliminated the need for Sound Blaster compatibility.

More than stereo

In the middle of the 1950's the stereo concept was developed. Music is recorded in two channels - right and left. During the last 20 years work has been done on expanding the stereo concept, to make 3-dimensional sounds. And this can be done easily. Even with only two speakers you can create a sensation of depth in space. Many sound cards have built-in 3D sound effects like Virtual Dolby. In this way games can be equipped with even more realistic sound effects.

Sound on the Internet

One of the funny features about Internet is all the music you can retrieve. Primarily it is electronic, synthetic music. If you do not like that variety, well then there is not much to get. I have just recently realized the possibilities, and I want to share my experiences with you here.

There are different kinds of music you can find on the net. I know about two of those:

- MIDI compositions, which are "real" pieces of music, written for playback with any sound card. MIDI is a standard in Windows 95, so any PC with a sound card can play these files.
- Koan music. This something entirely different. It is genuine computer music, with much more potential than the "flat" MIDI files. Koan requires the addition of a *plug-in* to your browser to enable playing the files. Koan music is written to designated sound cards. The Sound Blaster AWE is the best as far as I know. I just have an ordinary Sound Blaster 16, and there is also a lot of good Koan music for that.

Please write to me if you disagree or want to suggest additions to this page. I am not particularly knowledgeable in PC sound but it is an interesting issue.

Koan music

Koan is an electronic music standard. And this represents a fascinating technology developed by the British company SSEYO.

Koan is "live" music - it changes every time you play it. You can compare it with an aeolian harp, where the wind and thus the tone is different each time it is used.

"I too think it's possible that our grandchildren will look at us in wonder and say, 'You mean you used to listen to exactly the same thing over and over again? Brian Eno 1996

The Koan music consist of small files, which start a process in the PC where they work. There may be 8 hours of music in a 12 KB file! So it is not the music itself which is contained in the file. Rather the files contain some structures, frames if you wish, about a composition. These frames are activated in your PC's math processor. Then the music is generated within your PC, differently each time you play it.

- Koan music is written specifically for a certain sound card. So you must have either Sound Blaster 16, 32 or 64. or a few other makes. Here again is a good argument to stay with the SB sound cards. They will give the fewest problems.

The music is Internet suitable, since the files are small. I have found music in the category ambient, that is long electronic music sequences. They can be very quiet and meditative, but they can also be more rhythmic.

If you

- Like electronic music á la Tangerine Dream and Brian Eno
- Have a SB sound card and speakers in your PC system

then you ought to try some Koan software. It is really simple to install and requires only little space.

How do I?

I write this on three premises:

- You have a Sound Blaster sound card and speakers.
- You are on the Internet and use either Netscape or MS Explorer.
- You know how to download and extract (unzip) files.

My installation example is based on software for Sound Blaster16 and Netscape in 32 bit Windows 95 edition (Netscape Gold, version 3 or 4). It may sound complicated as I describe it. However it is really quite simple:

You want to install SSEYO software, so you can play the small SKP files with exciting music. First get the following: 32 bit Koan Software for Windows 95, and SB16. You have to find the file on SSEYO home page. New versions arrive all the time.

It is a self extracting Exe file about 300 KB big, which you place in some temporary folder. That file will be deleted after installation. Run the file (it is called knp1032.exe), which will install the necessary plug-in in Netscape.

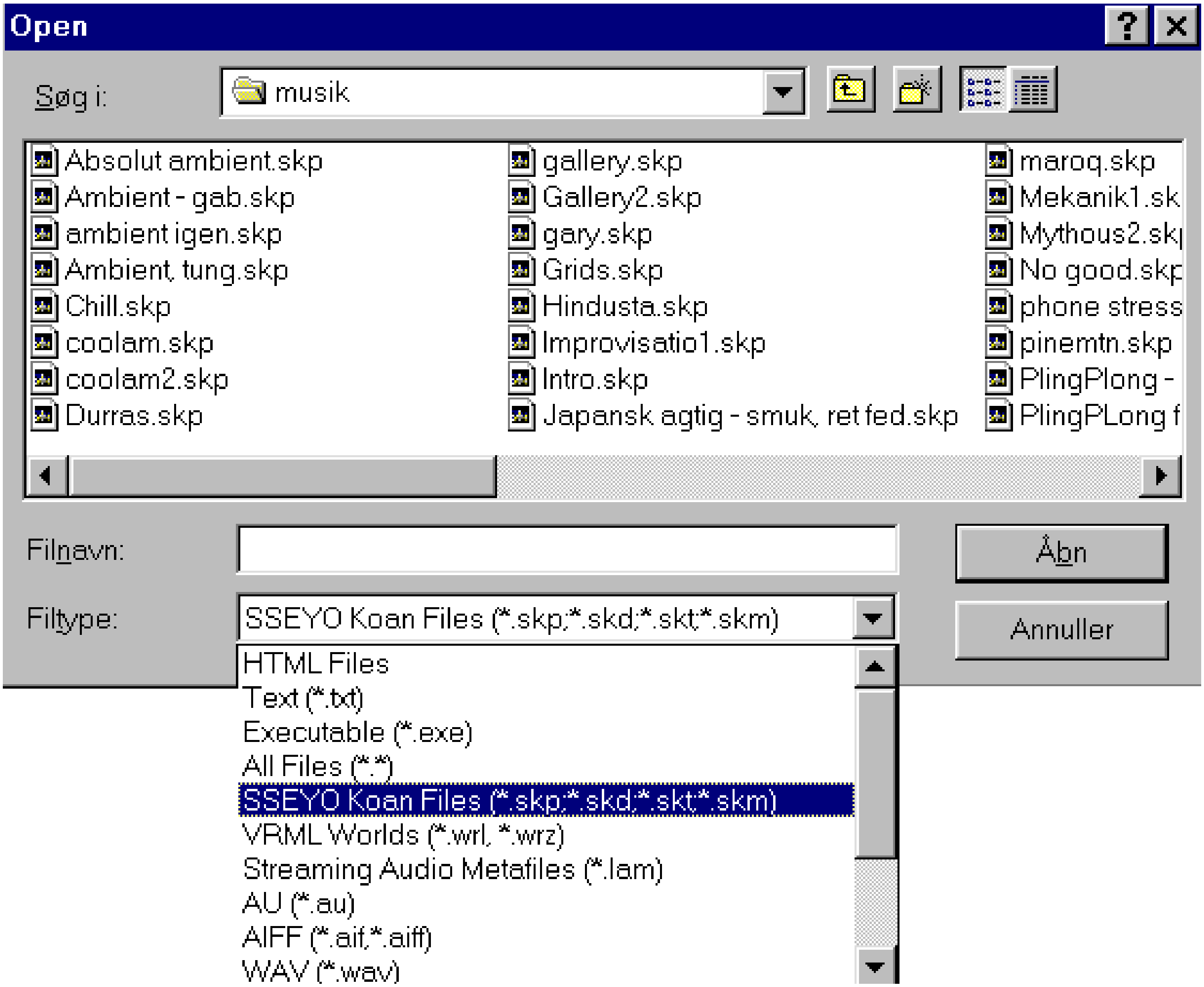
Now you can go on the net and for example retrieve the starter package on the same server, which includes some SKP files. Each of those represents hours of electronic music.

The best result is obtained with a AWE64 sound card and a pair of good speakers. I purchased a set of Altec Lansing ASC45. They are two tiny satellites with a heavy sub-woofer, and giving fantastic sound - that is Hi-Fi!

Play back of Koan music

Once you have installed Koan plug-In, Netscape can play the Koan files! It sounds backwards. You would think that the filter should work in Windows 95, but no - the music has to be played through an Internet browser. I use navigator for Koan (it works there) while I have chanced into Internet Explorer for surfing....

You save your Koan files in a folder (in my computer: **\web\music**). Then, in Netscape, you press Control-o (for *open*). Now you have to modify the file type, to activate the filter:



Now you just select the melody, and Netscape will play it. You need not be on the Internet, you just use the browser to play the music. The music can run in the background all day, while you do something else.

Links

SSEYO Koan home page where you can find a lot of information, including software to let you write your own Koan music.

There is a lot of MIDI music on the Internet. However, compared to Koan is is rather tame. There are rarely more than a few minutes of music in a MIDI file, and you soon get tired of the same pieces. The advantage of MIDI is that the file format is so standardized. If you have a sound card, no matter which, it will work.

Here is a link to a young person, who tries to create exciting music in the MIDI format. His page also include a lot of other interesting music links: Brazilian MIDI music.
