Graphics Cards

Introduction

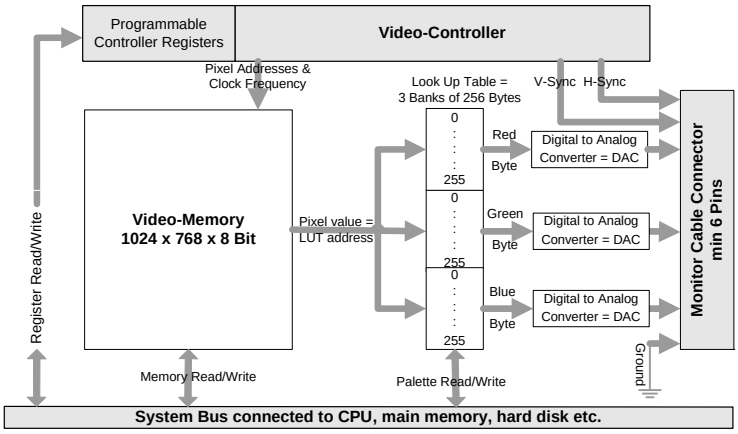
Graphics cards fill the pixels into the display. They obtain data from the main memory, the CPU and the periphery via a fast graphics bus (=input) and transform these data into a raster image that is transported to the raster display (=output). If there is a remote display you will need a monitor cable connecting the card and the display.
Just vector computers and computers without display (servers, robots etc) need no graphics card.

In most cases a graphics card is an expansion board plugged into a slot on the motherboard.
It carries its own CPU named Graphics Processing Unit GPU and its own memory chips = video memory.
The graphics card is an autonomous computer
who doesn't know that it works in parallel to a nearby CPU.
Cheap computers have cheap GPUs with no video memory.
They just obtain a fixed separated space of main memory that the CPU can't access anymore but even such GPUs are living their own lives with their own clock and address space.
Small mobile devices such as a navi, a mobile phone,
a digital camera don't have a dedicated graphics card but they have its main components somewhere hidden on the tiny main board.
These components are basically the same as real graphics cards have.
They still are autonomous from the rest of the computer even when the GPU is mounted some millimeters away from the CPU or when both are integrated in one single chip.
Summary:
Wherever there is a pixel, there is some sort of graphics card behind it.

Games, videos and Windows 7 need a lot of processing power. The addition of a dedicated GPU frees the CPU from graphics processing (green).
A computer is called a graphics computer, when its graphics card with GPU and video memory is more expensive than its motherboard with CPU and memory chips. In this sense, today's computers in tower cases are mostly graphics computers. Windows 7 and 3D computer games take that for granted.

See: http://en.wikipedia.org/wiki/Graphics_card
See: http://en.wikipedia.org/wiki/Graphics_processing_unit
See: Don Woligroski's Graphics Beginners' Guide, Part 1: Graphics Cards
See: Don Woligroski's Graphics Beginners' Guide, Part 2: Graphics Technology
See: Don Woligroski's Graphics Beginners' Guide, Part 3: Graphics Performance

8 Bit Graphics Card with LUT



This is the archetype of a graphics card without GPU. You cannot buy such a card anymore, but it is still hidden in any modern graphic card.
They boot in this mode and Windows uses it in its so called "Secured Mode".

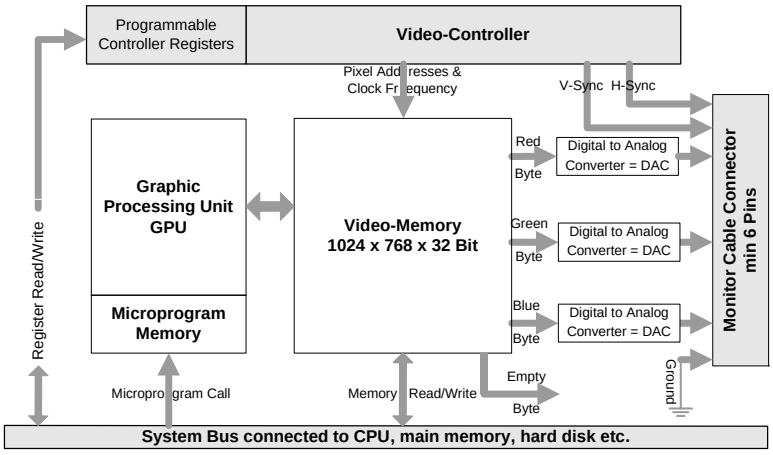
The **Video-Controller** is the central clock of the card. It produces the V-Sync and H-Sync signals that an analog raster display needs for vertical and horizontal direction of the electron beam. Since the video controller always knows the beam position, it knows, what pixel must be read next from the **Video Memory**. It therefore generates the memory address of the next pixel and triggers its reading. The programmer can change the resolution of 768 lines with 1024 points per line by setting the **Controller Registers**. The default resolution is very low: 600x480.

The **Video Memory** is depicted as a rectangle to suggest a pixel matrix. In reality it is a linearly addressed memory of ca. 1000x1000 = 1 Mio. bytes. Each one-byte-pixel can hold values between 0 and 255.
These values do not code a color. One-byte-pixel images have no color of their own. Color is coded in a separate 3x256 = 768 byte array called "**Palette**".
This palette is loaded in three 256-byte buffers called **Look-Up-Table = LUT**.
Any byte coming out from the video memory just serves as pointer in each of the three banks of the **LUT**.
Any pixel value points to a red, green and blue color value which are predefined by the palette.
Advantages:
1) Color coding is very flexible. It's no problem to simulate fast movements just by pushing a new palette into the LUT without touching the video memory. i.e. this is the base of all fast 2D-video games.
2) Economic use of video memory.
Disadvantages:
1) Any picture needs two data structures: a) the byte matrix ≈ 1 megabyte and b) the palette = 768 bytes.
2) The no. of simultaneously visible colors is limited to 256.
3) It's impossible to show videos.
The **System Bus** is not able to transport more than 5 pictures /sec to the **Video Memory**.

Look Up Table = LUT: Is a 768 (=3x256) byte SRAM containing three color hash tables = palette. Any pixel value serves as a logical color number which addresses three separate banks of the table (red, green and blue bank) that feed the three Digital to Analog Converters = DACs.
Palette: Hash values in the LUT. A palette can code 256 different colors. Whenever other colors are needed, you must load another palette into the LUT which defines another subset of 256 colors out of the total scope of 256^3 possible colors. The header of a 8-bit image normally carries a suitable palette.
Images with true color (24 bit per pixel) don't need a LUT with palette. The three bytes of the pixel are fed directly into the DACs, bypassing the LUT, which is out of job.
RAMDAC: A microchip on the graphics card containing LUT and 3 DACs. It connects video memory with the red, green, blue connector pins of the monitor cable.
It combines a small static RAM (SRAM) containing a color palette table with three digital-to-analog converters (DACs) that change digital pixel data into analog signals that are sent to the electron guns, one for each primary color - red, green, and blue.

Output:
At the back side of the PC you will see one or two connectors where the display cable plugs into.
Usually there is D-Sub 15 pin connector and a DVI 24 pin connector.
1. Output: Analog VGA via D-Sub connector: for cathode ray tube monitors.
2. Output: Digital Video Interface = DVI: for flat panel monitors (not shown in the picture).
The 3 parallel bytes coming out from the LUT bypass the DACs and are fed directly into DVI.
The problem of DVI is the transport of the enormous pixel data rates over long distances (> 2 m).
Old graphic cards and old flat panel displays don't have DVI. They have to convert their data twice:
1. Graphic card: digital to analog conversion via DACs.
2. Transport of an analog signal through the video cable.
3. Flat panel display: analog to digital conversion via ADCs to obtain x, y-coordinates that address a pixel.
Such double DA+AD-conversions reduce picture quality.
Notice: Modern graphics cards contain such a 8-bit graphics card as a subset.
The 8 bit mode is active at any computer start at the beginning of the boot sequence (and in "safe mode") when text and logos appear on a black background.

32 Bit Graphics Card with GPU



The next generation of graphics cards has a pixel depth of 32 bit and the RGBA-pixel-format: One byte for Red, one for Green, one for Blue and one for Alpha = transparency. This format allows "true color" images where each pixel carries its own color information.

Such pictures need no palette nor LUT. They need a 4 times bigger video memory and a GPU.
The GPU is irredeemable while the system bus is unable to furnish new RGBA-images at a sufficient rate.
The GPU renders the content of the video memory locally.

Advantages of a GPU:
1) disburdens the CPU from producing images
2) disburdens the system bus from transporting images

Disadvantages of a GPU:
1) the graphics card is a costly second computer
2) the operating system must provide a graphics library of microprograms = Drivers
3) the GPU just produces artificial images.
Real-world-images still need the bus.

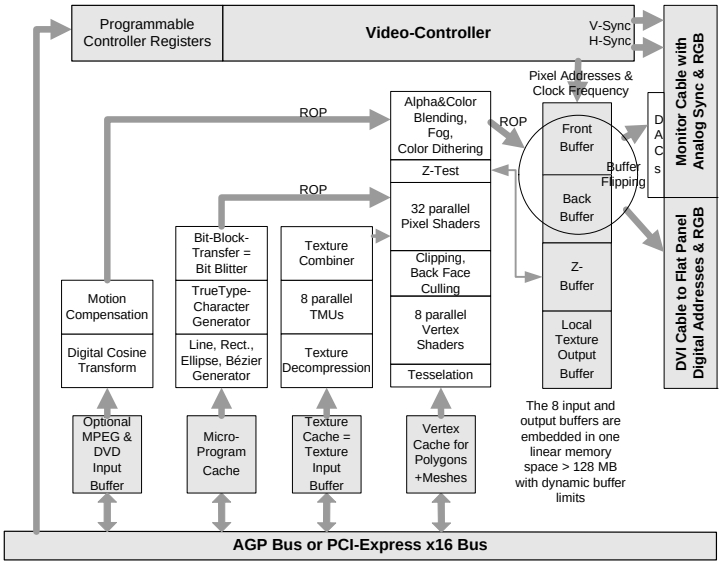
Differences of a 32 bit graphics card compared with a 8 bit graphics card:
1. Needs quadruple video memory and quadruple bus bandwidth.
2. No LUT. Cannot display 8 bit images with palette.
3. Specialized on true color = red, green, blue (RGB) bytes plus an empty byte.
4. Contains its own CPU = GPU, because the system bus is too slow.

Samples of GPU-tasks:

GPU should render	Microprogram Call	Parameters
Text	DrawString	ASCII-string, font-size; normal- or bold- or italic-indicator
Line	DrawLine	two 2D-coordinates: x0,y0,x1,y1; Pen-object
Rectangle	DrawRect	one 2D-coordinate: x0,y0; width; height; Brush-object
Ellipse	DrawEllipse	same as Rectangle
Scroll down	ScrollDown	no. of lines, ASCII-string of the new lines to be inserted at the top

Such a 32-bit graphics card doesn't solve the slow-system-bus-problem for real world images and videos.
To solve the problem, Intel proposed a separated system bus specially for graphics cards.
1) AGP = Accelerated Graphics Port provides a fast data transfer from CPU, main memory and harddisk to the graphics card. This bus allows much higher transfer rates as the common 66-MHz-PCI-Bus. It exists in 4 velocities: AGP, AGP2x, AGP4x, AGP8x, the latter has 2,1 GByte/sec in direction of the graphics card, but only 264 MByte/sec in the opposite direction.
2) PCI-Express = Peripheral Component Interconnect Express replaces AGP since 2004. Its fastest version PCI Express x16 = PCI Express for Graphics = PEG allows 3.7 GByte/sec in both directions. See: **PCI Express**

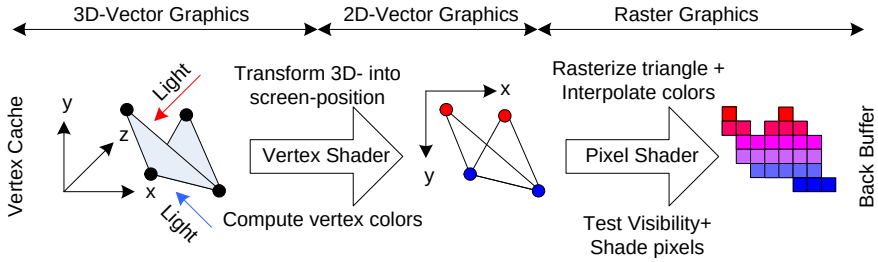
64/128/256/512 Bit Graphics Card with 8 Buffers



Modern GPUs host a network of specialized processors (up to 200) on a single chip. The numbers 64/128/256/512 do not refer any longer to the pixel depth of Video Memory but to the no. of parallel wires of the internal graphics bus. 64 means a bus width of 64/32 = 2 integers, 512 means a bus width 0f 512/32 = 16 integers which are processed at once by 16 parallel shaders.
A selection of some important processors appear as bright adjacent boxes in the mid of the image. The most important are called Shaders.

The Video Memory is soft-divided into 8 segments with moving borders:
Output Buffers: (dark squares at right)
1) Front Buffer contains the current display content.
2) Back receives the new display content.
3) Z Buffer contains the pixel-depth-values of the Back Buffer.
4) Local Texture Output Buffer contains sprites and repeating textures.
Input Caches or Input Buffers: (dark squares above the bus)
1) Vertex Cache contains the mesh's Vertex- and Index-structures.
2) Texture Cache contains all images.
3) Micro-Program Cache contains the graphics library.
4) MPEG & DVD Input Buffer contain compressed video streams.

The GPU has two major tasks:
1) transform 3D-Vector-Graphics to 2D-Vector-Graphics = projection
2) transform 2D-Vector-Graphics to Raster Graphics = rendering



Shaders = Shader Units = Shader Processors:
Vertex Shader (formerly Transform&Light Engine = T&L Engine) is a programmable pipeline of processors specialized in vector graphics. Up to 8 of such parallel pipelines scroll, zoom, rotate vertices and normal vectors. See: **Wikipedia**
Pixel Shader is a programmable pipeline of a texture processor and two arithmetic & logic units (ALUs) performing multiply/add instructions (MADD) specialized in raster graphics. Up to 8 groups of 4 = 32 parallel Pixel Shaders stretch and combine textures (see Texture Filtering below). See: **Wikipedia**
Unified Shader: The upcoming DirectX 10 treats both vertex- and pixel-shaders as a unified structure. This new specification can be found inside the Microsoft Xbox 360.
Vector Pipeline: Flow of vertex data through the vertex shaders. See: **Wikipedia**
Pixel Pipeline: Flow of pixel data through the pixel shaders. In new architectures the name pixel pipeline lost its meaning as a pixel shaders is no longer attached to a single TMU. Sample: 12 pixel shaders can be combined with 4 TMUs.
Texture Mapping Unit = TMU: TMUs address, filter textures and apply texture operations to pixels. GPUs contain several parallel TMUs. See: **Wikipedia**
Raster Operator Unit = Render Output Unit = ROP: ROPs read and write pixel data to the Video Memory. The speed at which this is done is known as the fill rate. See: **Wikipedia**
Memory Bus: Width can range from 8 bits to 512 bits. As the bus width increases, so does the amount of data that the bus can carry per cycle.

The Video Memory is soft-divided into 8 segments with moving borders:
4 output Buffers: Front, Back, Z, Local Texture Output
4 input cache Buffers: Vertex, Texture, Micro-Programs, MPEG.
Output Buffers contain raster matrices of 32 bpp.
Front-, Back- and Z-Buffer always have identical size and structure. They can replace each other.
The Front-Buffer contains the currently visible image and is constantly addressed by the video controller to feed the DVI and/or the DACs. In the meantime the GPU writes the next image into the Back-Buffer.
Frame Buffer Flipping = Front- to Back-Buffer Flipping:
When the Back-Buffer is completely filled, back and front exchange their roles. The old Back-Buffer now plays the role of Front-Buffer and the old Front-Buffer takes over the role of Back-Buffer. The old content of the ancient Front Buffer = new Back Buffer will now be overridden by new pixels coming from the GPU.
This Buffer exchange is called "Buffer-Flipping" and is performed during the refresh (V-Sync on) of the monitor. Buffer-Flipping separates the pixel writing from 1. GPU into the Back-Buffer from 2. pixel writing from Front-Buffer into the monitor. This separation together with the fast exchange of roles during the short dark period between two monitor images prevents from flicker artifacts.
Even faster results are obtained by using one Front-Buffer together with two Back-Buffers which prevent from waiting for V-Sync on.
The above schematic image of a 64/128/256/512 Bit Graphics Card with 8 Buffers shows one Front- and one Back-Buffer at fixed positions with a connecting circle carrying the name "Buffer Flipping". The Buffers exchange their functions without any copying. Input and output connections switch symmetrically at any flip.
Z-Buffer: See: **Wikipedia**
The Z-Buffer (being part of the overall Video Memory) contains no color but depth information (camera distance of any pixel using the data type = long integer). Before the graphics processor writes any color into the Back-Buffer, it compares the new z-value to the existing one.
A pixel is written to the Back-Buffer if the old z-value is greater than the new one because in this case the new element is in front of the old one.
Z-matrix must be initialized with the highest possible Z-values. Z-Test:

```
if ( [z]-value of Pixel[x][y][z] < Z[x][y] )
{ copy PixelColor[x][y] to Back-Buffer[x][y];
  copy [z]-value to Z[x][y];
} else forget Pixel[x][y][z];
```

Older graphics cards used a 16 bpp Z-Buffer. This works fine as long as the rounding errors do not accumulate. Such rounding errors tend to add up to severe depth artifacts. This is the reason why 32 bpp is today's standard of Z-Buffer depth.

Local Texture Output Buffer:

A texture is a 2D image matrix of arbitrary size that contains color pixels (= texels) to be copied into the Back-Buffer. Multi-Texturing is the presence of multi-layer textures in 8 different byte planes. Specialized "texture combiner processors" inside the GPU add, subtract, multiply, stretch, interpolate, filter two or more byte planes to complicated special effects such as " Lighting-Mapping, Bump-Mapping, Dot-Product3-Bump-Mapping, MIP-Mapping, Environment-Mapping, Alpha-Blending".

Intel announced its entrance in the market of high end graphics processors using a new massive parallel architecture of more than 32 x86-Pentium processors: **Larrabee**. Unlike NVIDIA and AMD-ATI Intel doesn't align special fixed-function processors along a pipeline as described in this paragraph. Intel tries to solve the problem with cheap and energy-efficient parallel >= 32 multi-purpose processors and a sophisticated C-compiler which splits the DirectX data stream in parallel tasks.

Technical Terms

Naming Convention: A graphics card name carries the width of its Memory Bus.
1. A 8 bit graphics card has a pixel depth (= **bit per pixel = bpp**) and a Memory Bus width of 8 bit.
2: A 32 bit graphics card has 24 bpp or 32 bpp and a 32 bit Memory Bus.
In case of 32 bpp, the last byte of a pixel contains nothing.
3: A 64/128/256/512 bit graphics card has 32 bpp, but carries 2/4/8/16 adjacent pixels at once.
The processor communicates with its video memory using a 64/128/256/512 bit Memory Bus.
Thus any bus cycle transports 8/16/32/64 bytes in parallel and all video memory addresses must be divisible by 2/4/16/32.

This parallelism complicates graphics programming, because it is ineffective to address a single pixel.
Any fast algorithm must try to work with 2/4/16/32 adjacent pixels.

Video Memory: Normally consists of 4 identical DDR-RAM chips 64 MByte each = 256 MByte in close vicinity to the GPU (sometimes hidden under the GPU cooler). This memory is linearly addressed without any segmentation. The logical segmentation into 4 output buffers and 4 input buffers is floating when the screen resolution changes.
Sample for a 1600x1200 screen: Front- Back- and Z-buffers need 1600x1200x4 = 7,6800,000 bytes each ≈ 23 MB altogether ≈ 10% of overall Video memory.

Pixel Fill Rate: is the total number of pixels the card can output and is calculated as the number of raster operations (ROPs) multiplied by the clock frequency.

Texture Fill Rate: is the number of pixel pipelines multiplied by the clock frequency.

High Dynamic Range Lighting = HDR: Contrast enhancer (darker darks and brighter lights), while at the same time increasing the amount of lighting detail displayed in both the dark and bright areas. See: **Wikipedia**

Anti-Aliasing = AA: Reduces jagged or blocky patterns and stair-likeness of angular edges of the raster. Anti-aliasing calculations use a fair amount of graphics processor power → drop in frame rates. See: **Wikipedia**

Texture Filtering: Reduces blockiness of highly enlarged textures. See: **Wikipedia**
1. bilinear filter: Simple color interpolation with the 4 neighbors of a texel.
2. anisotropic filter: Filter with trapezoid shape according to view angle.

Texture Set: Collection of textures that fit into the Video Memory. Game programmers have to provide different texture sets of different sizes because they don't know whether they target a 128, 256, 512 or 1024 MB Video Memory. Texture Sets that are too big for a given graphics card have to be stored in slower system RAM, or even on the hard disk which heavily taxes performance.

Overdraw: In 3D-scenes very often polygons lay in front of other polygons. There is no effective way to know which polygons are hidden. The useless rendering of hidden polygons is called "Overdraw". In 3D-games the Overdraw-factor is about 3, that means that 3 pixels have to be computed and written to the output buffers in order to see one finally. This Overdraw-factor has to be taken into account for the necessary performance of the processor and the bandwidth of memory, it heavily limits performance.
A modern method to reduce Overdraw is called Z-Culling, a feedback loop between the Z-comparison of the Pixel Shaders and the Clipping and Back Face Culling unit. Whenever the Z-comparison rejects a pixel, the triangle is tested whether it is completely covered or not. Z-Culling reduces Overdraw by ca. 50%.
The ultimate challenge of graphics programming is a 100% elimination of Overdraw by "Deferred Rendering". Deferred Rendering requires tough programming and more memory space to store "Display Lists" = data structures representing all current 3D-polygones. These display lists enable to determine in advance the pixels in front of all others and to reduce Overdraw to zero.
The graphics chip inside Sony PlayStation 2 has this architecture.

More **Technical Terms** are explained here: **english** or **deutsch**.

Chipset GPU and Graphics-Card-GPU

There are two types of GPUs:
a) Chipset GPU = Mainboard GPU:
Many notebooks, barebones and in office desktops just have low-cost GPUs (less than 10$) integrated within the chipset on the motherboard. Such cheap GPUs mostly are stripped down versions of the much faster (and more expensive) Desktop GPUs. The differences are:
a1) low clock frequency → low power consumption
a2) just a few pixel pipelines → low power consumption
a3) no private graphics memory → They steal a part of main memory instead.
Samples: Intel G950, AMD-ATI Radeon Xpress 1250, Nvidia GeForce 7050
b) GPU mounted on a graphics card = Desktop GPU for games and multimedia:
b1) NVIDIA (Santa Clara, Cal) → **www.nvidia.com/page/home.html**
b2) AMD-ATI (Ontario, Can) → **http://ati.amd.com/**
b3) Intel (Santa Clara, Cal) → **Larrabee**
Comparison of recent graphics cards: **www.techarp.com/showarticle.aspx?artno=95&pgno=3**

The strategic fight between Intel and Nvidia

Intels business has always been driven by new operating systems and new applications needing more computing power. But the public begins to loose interest in multi-core CPUs that are underworked. Intel's answer is to expand its multi-core CPUs into hybrid CPU/ GPUs → **Larrabee**.
Nvidia counter attacks with its **CUDA** programming interface that uses the GPU as processor for all sorts of parallel computations expanding Nvidia GPUs into hybrid GPU/CPUs. See: Interview mit Jen-Hsun Huang.

Intel is in a better strategic position because Microsoft's and Apple's operating systems run on Intel platforms and neither Microsoft nor Apple plans to rewrite them for Nvidia GPUs.
Both Larrabee and CUDA are aimed to bridge the gap between CPU and GPU and to unify both worlds into one massive parallel processor.

Abbreviations

EGA	Enhanced Graphics Adapter	IBMs 1985 standard for graphic board resolution=640 x 350 and color depth=4 bit	Wikipedia
VGA	Video Graphics Array	IBMs 1986 standard = 640 x 480 and color depths=4 bit to 8 bit	Wikipedia
SVGA	Super VGA	800 x 600 x 8 bit	Wikipedia
XGA	Extended Graphics Adapter	IBMs 1990 standard = 1024 x 768	Wikipedia
SXGA	Super Extended Graphics Adapter	1280 x 1024	Wikipedia
UXGA	Ultra Extended Graphics Adapter	1600 x 1200	Wikipedia
VESA	Video Electronic Standard Association	Standards for hardware, timing, commands of display connections	www.vesa
LUT	Look Up Table	768 = 3x256 bytes of static RAM (=SRAM) containing the color palette	Wikipedia
DDR-SDRAM	Double Data Rate Synchronous Dynamic RAM	Memory-IC (=Integrated Circuit) with 16x or 32x bus width for graphics cards 2004: DDR-SDRAM 2,5 Volt, up to 400 MHz → on standard graphics cards 2005: DDR2-SDRAM 1,8 Volt, up to 533 MHz → on higher graphics cards 2006: DDR2-SDRAM up to 667 MHz 2007: DDR2-SDRAM up to 800 MHz 2008: DDR3-SDRAM 1,5 Volt, up to 1600 MHz	Wikipedia
eDRAM	embedded Dynamic RAM	Dynamic RAM inside the graphics processor. Graphic memory is directly on the processor chip. No separate memory chips and no bus structure necessary on the graphics card → fastest access to graphics data.	NEC
RAMDAC	Random Access Memory plus Digital to Analog Converters	Chip on graphics cards containing four components: small static RAM (containing the LUT) plus three digital-to-analog converters (DACs)	Wikipedia
GPU	Graphics Processing Unit	Mega-chip on graphics card containing multiple specialized processors	Wikipedia
T&L	Transform and Light Engine	Vertex Shader sub-unit of the GPU executing prefabricated firmware	Wikipedia
VGA D-Sub 15		This 15 pin analog monitor connector was the standard before DVI and is still very common.	Wikipedia
DVI	Digital Video Interface 	Connector- and cable-layout for digital pixel data transfer to flat panel displays. DVI avoids superfluous and noxious DA-conversion inside the graphics card and AD-back-conversion inside the flat panel display. It carries 2D-adresses (x,y) and digital RGB data instead of analog Syncs and analog RGB signals. There are two different DVI connector layouts: DVI-D with 24 purely digital pins and DVI-I with additionally 5 analog pins for CRTs.	Wikipedia
HDMI	High Definition Multimedia Interface 	HDMI is the coming standard carrying both DVI and 8 uncompressed digital 24-bit-audio-channels via one 19-pin-cable. To view HD-video on a PC requires both an HDMI video card as well as an HDMI-enabled flat panel monitor.	Wikipedia
AGP	Advanced Grapics Port	AGP is a high-bandwidth interface designed specifically for graphics cards with direct read/write capabilities with the system memory, de-multiplexing or simplification in the organization and transfer of data, and increase clock speeds. AGP has gone through three major revisions, with the newest being AGP 8x at 2,1 GB/s.	Wikipedia
PCIe	Peripheral Component Interconnect Express	PCI Express is a serial interface which runs with few connections. Different from parallel buses, the total bandwidth is available for every bus slot. PCI Express x16 (16 links) offers a bandwidth of 4 GB/s up and down or 8 GB/s total.	Wikipedia