Computer Graphic Cards

Evaggelos Thanos

SOME GRAPHIC CARDS

AMD	ASUS	GEFORCE	PALIT
			

Table of Contents

Overview
This technical report will discuss how a computer's graphic card functions. The document starts with a broad overview of the graphics card, and progressively gets more into depth about the technical issues of each component of the graphics card, and ends with a future outlook on graphic cards.

Graphic Card Summary
A computer graphics card is an additional peripheral that enhances the performance of rendered graphics. The card mainly consists of a *graphical processing unit* (GPU), memory, and a digital/analog converter and connections to and from the graphics card. Most applications on a computer requires some type of graphics to be displayed. The information is processed from the application to the *central processing unit*(CPU) which is then sent to the specialized graphics accelerator for quicker processing. After undergoing the transformation of computer code, the data is then sent to a monitor to be displayed.

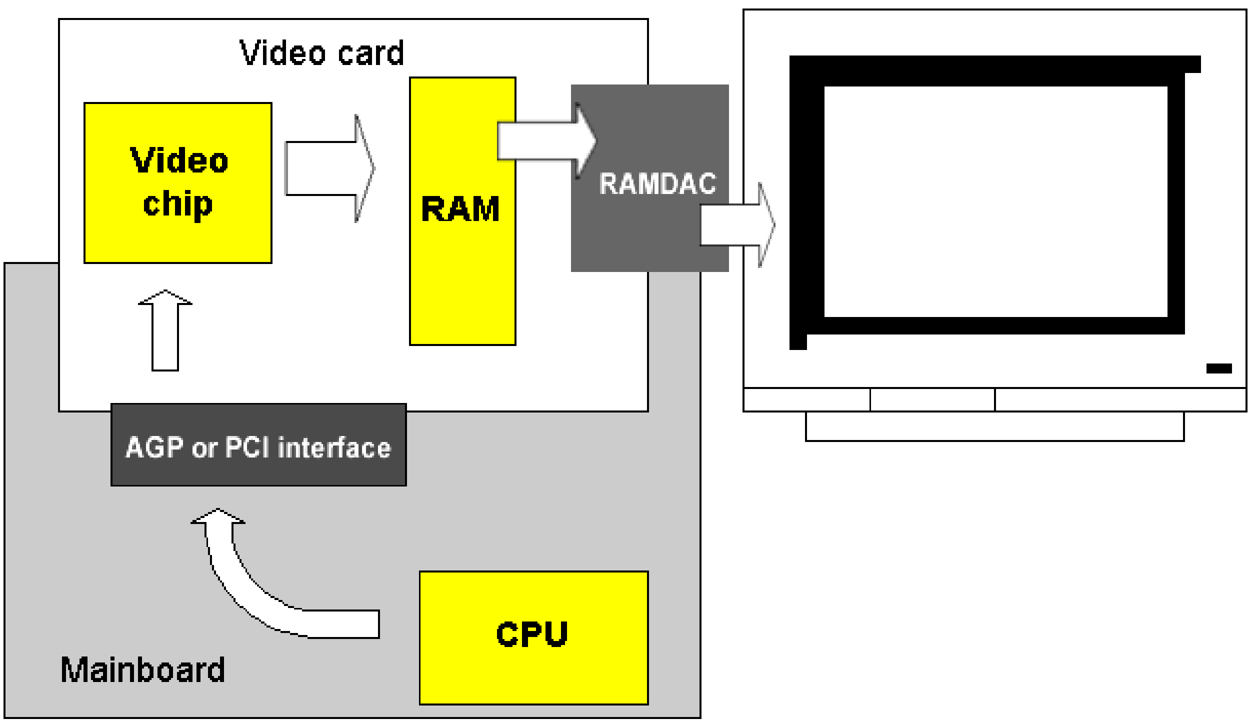


Figure 1. General Flow Diagram of Graphics Card

Background on Computers
Originally, computers were once all text based and did not come with a *Graphical User Interface* (GUI) that consumers have grown accustomed to. As computer technology has advanced, graphics has become an important part in the way humans interact with computers and the

demand for graphic oriented computing has called for the need of specialized graphical processing. This is when the age of graphical accelerated cards began.

Most computer chips are produced as silicon chips. A silicon chip is composed of thousands of transistors, which are in turn composed of three layers of conducting material, mainly silicon, forming a "sandwich" design. The layer might contain either a positive or negative type of silicon. With a difference in electrical charge, and allowance to accept and receive electrons, transistors are used as switches. Current cannot pass through a transistor because of the diode effect. Diodes are devices that block current going in one direction while allowing the opposite direction of the current to flow. This is very important as it can keep sensitive electronics safe from a reverse charge.

Design

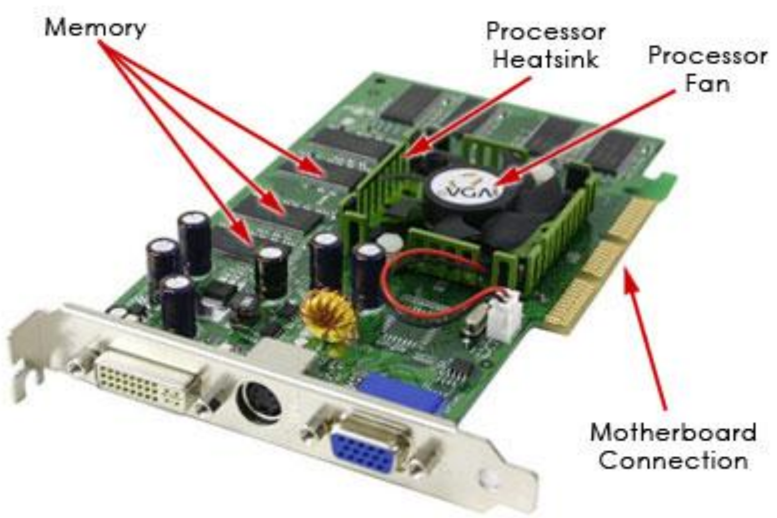


Figure 2. Basic Layout of Graphics Card

Graphical Processing Unit

The GPU is a processor that is specially designed for computer graphics. The processor accelerates graphical processing while taking the CPU's workload, which is used to process every instruction code for the computer. Since graphical processing units are devoted to processing graphics, the processing of advanced graphical algorithms and coding are accelerated compared to the regular CPU.

Pipelined Architecture

A pipelined architecture is the standard procedure for processors as it breaks down a large task into smaller individual grouped tasks. When a set of instructions are transferred to the GPU the GPU then breaks up the instructions and sends the broken up instructions to other areas of the graphics card specifically designed for decoding and completing a set of instructions. These pathways are called stages. The more stages the graphics card has, the faster it can process information as the information can be broken down into smaller pieces while many stages work on a difficult instruction.

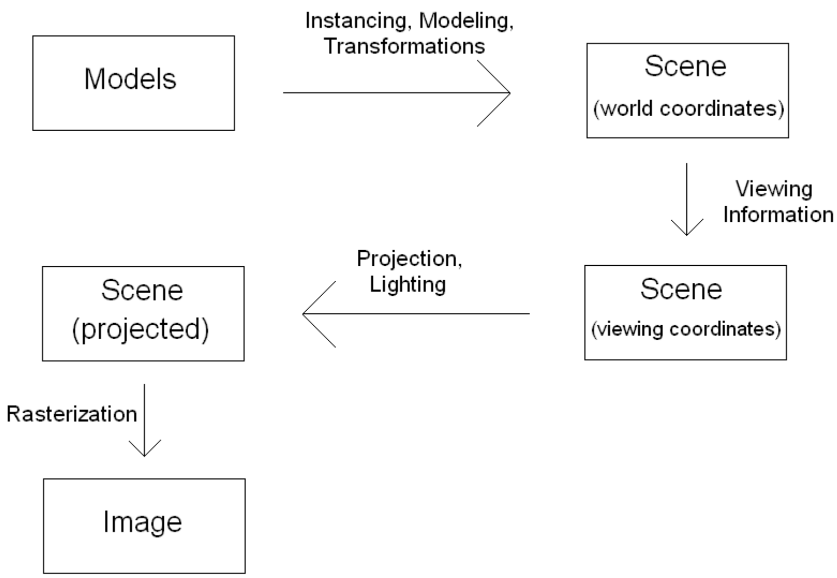


Figure 3. General flow diagram of graphical pipeline.

Pipeline Process : Stage 1

The pipeline process starts with an "Application/Scene" stage, also known as the workload-reduction trick. This stage is devoted to deciding which particular object will be rendered in the *three dimensional*(3D) environment. The way that 3D environments are created is through a Cartesian Coordinate Systems (an x, y, and z axis) in which objects are placed to create a scene. Scenes can have multiple angles to view them, in which they are created by reference points (cameras). The view space is determined by objects and angles depending on how the creator programmed the scene. The first process of the graphics pipeline is to only render and produce the images of the view space and to skip over unnecessary instructions that will not be displayed even if it was processed. This system allows the graphics card to produce scenes and graphics efficiently.

There are also factors that involve in processing the image efficiently, mainly the *Level of Detail* (LOD). An objects distance to the reference point of the view camera has an effect on the object's LOD. Some objects are assigned multiple resolution settings (the quality of image that is determined by the number of triangles that composes the object) in which the closer object might receive a higher resolution, while the same object farther away might be at a lower resolution to reduce the workload of the graphics card.

Pipeline Process : Stage 2

The second stage involves the scene's geometry. Objects mainly get moved from frame to frame to give an illusion that the object is moving in a real time setting. Objects can both be moved and manipulated in a scene depending on the application in which it is running. This manipulation of objects is generally called transformation. The objects can be stretched, skewed, moved or moved about an axis, or scaled differently. It is in the second stage of processing that the objects in the environment are altered.

Geometric lightning also occurs in the second stage after the objects are in their proper place, and once the figures receive their shape through the geometric transform process. Different types of lightning, depending on the application being run, will be processed to give the graphics a realistic appearance.

After the lighting is calculated the scene needs to get rid of unnecessary triangles that are only partially shown through the view space. This process is similar to the process which occurs in the first stage and also includes the process of "clipping." "Clipping is the operation to discard only the parts of triangles that in some way partially or fully fall outside the view volume" (http://www.extremetech.com/article2/0,1558,464440,00.asp).

Pipeline Process : Stage 3

The third stage implies an algorithm called the *digital differential analyzer* (DDA) which calculates the position of each part of all the triangles, and determines if the triangles are connected to other triangles. This process is done by computing the slope of each triangle's edge in hope to improve the quality of the image being produced and by allowing more detailed information to be assigned to the triangles Sometimes when two triangles are touching, or even overlapping each other, a rough pixel "stair-step" occurs in which the edges between the two triangles create non realistic images of edges extruded surfaces that would not normally occur.

Another thing that occurs in the triangle setup phase is the assignment of color and depth values for each pixel. Since the edges of the triangles were calculated, the color and depth values may be interpolated using each triangle's vertex vales of color and depth. Along with the color and depth, the texture coordinates of each pixel is also interpolated in which they will be processed in the fourth stage.

Pipeline Process : Stage 4

The last stage of the pipeline is considered the rendering / rasterization stage. "To fill the frame buffer the drawing primitives are subdivided into pixels, a process known as scan-conversion or rasterization" (Schneider, Benyt-Olaf, pg 245). After all the processes of computing location, color, geometric values, etc., this last stage puts it all together and produces the 3D environment onto a 2D screen.

After the triangle setup is completed in the third stage, the next step is to provide shading values. Shading values are similar to the color and depth values contained in stage 3 and add the finishing touches on the scene. There are three common shading methods: Flat, Gouraud, and Phong shading.

- Flat Shading - operates per triangle and provides a quick render of the scene that does not involve extensive computations. This type of shading does not produce a high quality image.

- Gouraud Shading - operates per vertex of the triangles. Compared to flat shading, Gouraud shading produces a higher quality image while sacrificing render speed. Because of Gouraud shading takes the lighting values of each vertex of the triangle and interpolates the values across the surface of the triangle, the object will appear smother and not as rigid as flat shading.

- Phong Shading - operates per pixel and is the most computation demanding shading process compared to flat and Gouraud shading. Phong shading incorporates the Gouraud Shading idea of taking the average shading of the vertices and also implies its own process that includes other triangle's shading as well. This makes the object blend together easier for more complex designs and results in a higher quality image making it more realistic.

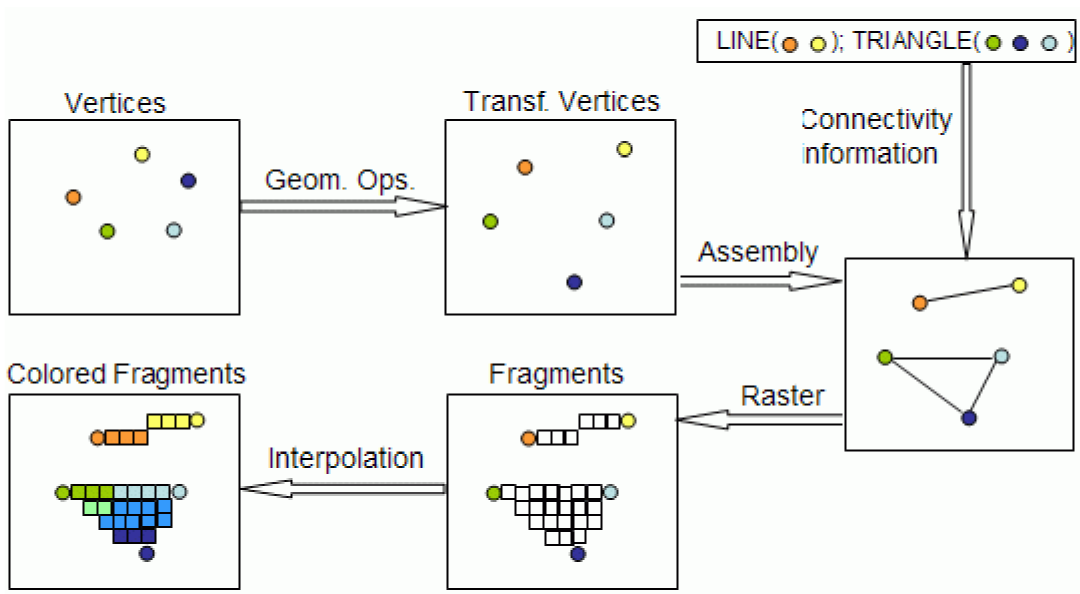


Figure 4. General flow chart of Gouraud/Phong shading

Memory

Random Access Memory (RAM) assists the graphics card process information. RAM is composed of transistors and memory cells that are arranged in a row and column grid that allows date to be stored and accessed quickly. New technology yielded the *Double Date-Rate Synchronous Dynamic Random Access Memory* (DDR SDRAM, commonly referred to as DDR), and the DDR2 RAM modules. These types of RAM modules are economically beneficial as they have higher efficiency, cost less, and have a higher potential for improvement.

The memory cells of the RAM can read either the number 1 or 0, which changes due to the capacitor change in gain or loss of electrons. Dynamic memory has a slight flaw, the capacitors have a natural electron leak and are drained of electrons every few milliseconds. To solve this problem, the CPU or the memory controller has to recharge all of the capacitors holding a 1 before they discharge. That is why RAM refreshes thousands of times per second. Memory cells have their own special support infrastructure of circuits that enables it to identify each row and column in the memory cell, keeping track of the refresh sequence, reading and restoring signals from a cell, and telling a cell if it should take a charge or not.

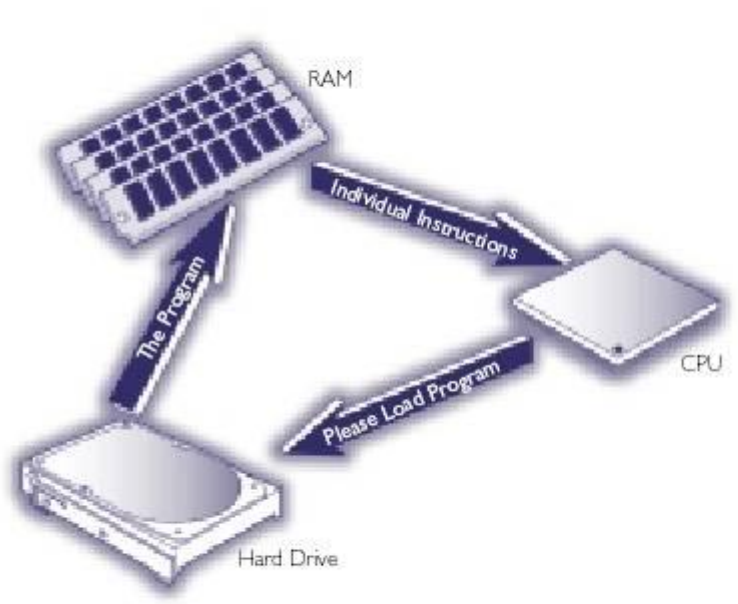


Figure 5: General flow of information exchanges by RAM

The memory of the graphics card (VRAM) is controlled by the GPU that stores data in specialized video memory storage space. This storage space operates by the use of common storage space but is especially reserved for graphical processing. The memory operates along with the GPU to produce quick instructions and processing that the graphics card can accomplish. VRAM is necessary to keep the entire screen image in memory. The CPU sends instructions to the video card which undergoes the graphical process and eventually is able to be displayed on a screen.

Connections

Graphics Card to Motherboard

- Computer Bus Connector - the graphics card itself is connected directly to the motherboard either through an Accelerated Graphics Port (AGP) or a Peripheral Component Interconnect Express (PCIe) slot.

Recently PCIe has replaced the AGP in becoming the quickest method to transfer information between a additional device and the computer. “A connection between a PCIe device and the system is known as a ‘link’ and this link is built around a dedicated, bi-directional, serial (1-bit), point-to-point connection known as a ‘lane’” (www.matrox.com). This allows the GPU and CPU to interact with one another at high speeds to process different instructions. Because of the PCIe high bandwidth there can be up to 32 “lanes.”

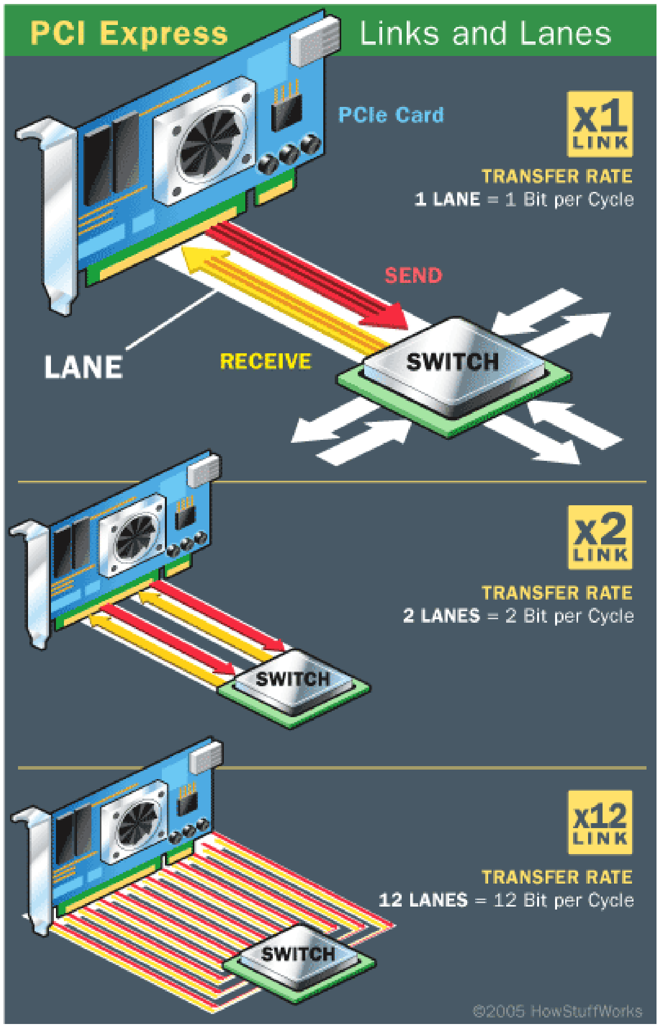


Figure 6: Overview of PCIe capabilities

- Graphics BIOS – (basic input/output system) performs diagnostics tests to make sure graphics card is fully functioning.

The graphics card has it's own BIOS that is similar to the motherboard's BIOS in which it prepares the device so that it may be used. Since the added complexity of graphic cards, a separate BIOS on the card itself has become a new standard. While not being used in the graphical process, the BIOS serves as a safety for the user and computer.

Digital-to-Analog Converter

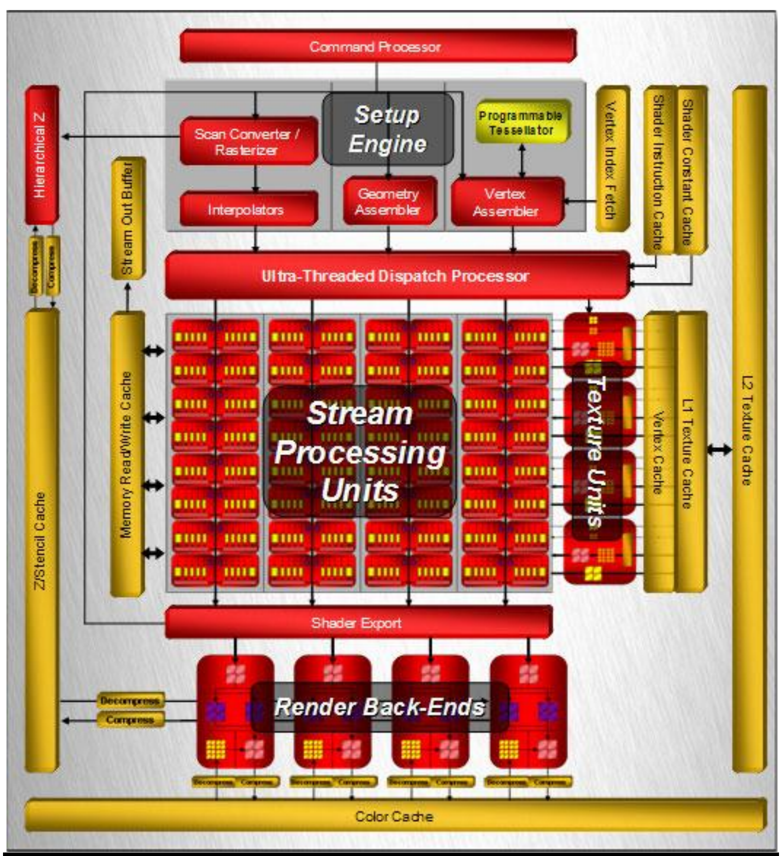
- *RAMDAC* - the *Random Access Memory Digital-to-Analog Converter* (RAMDAC) coverts the graphics card's digital memory into analog memory.

Digital-to-Analog Converters (DAC) are common in electronics and applies in the same way to the graphics card. The RAMDAC makes it possible to transform the digital data processed by the graphics card to be displayed on a monitor. After the graphics has been fully processed through the graphical pipeline and is waiting to be displayed, the information is sent to the

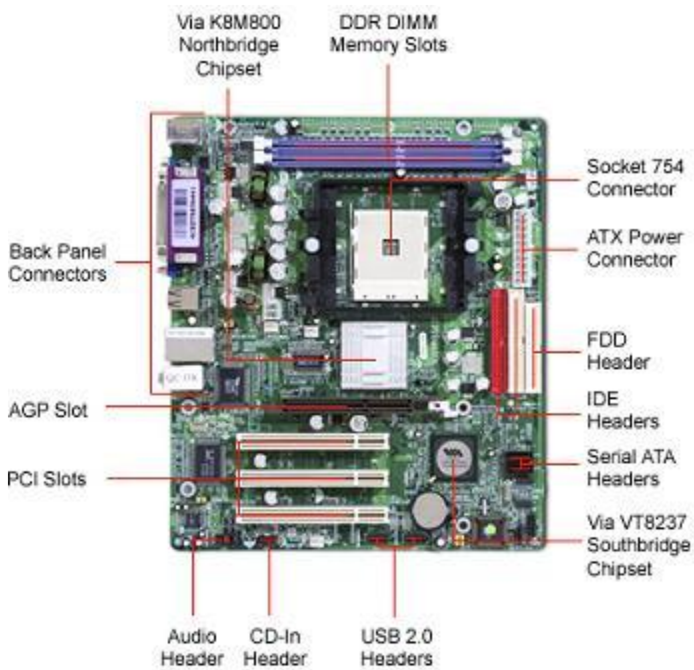
RAMDAC which takes the computer code and transforms the code into a different voltage level. The analog signal is composed of the red, green, and blue signals and is sent to a monitor via a video cable. This processed signal shows up on the computer monitor as the pictures and graphics.

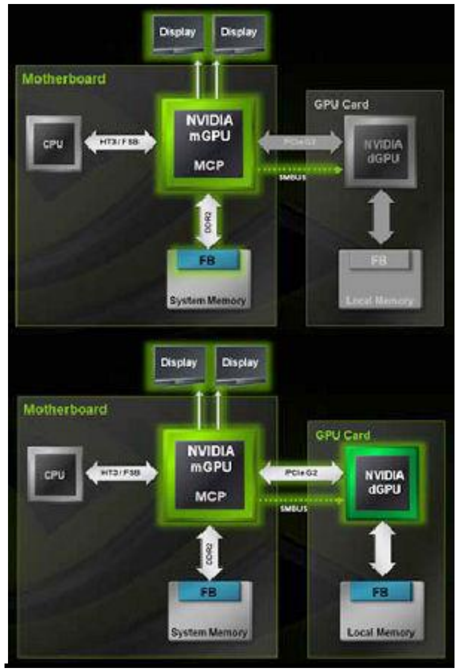
The Future of Graphic Cards

As applications demand more processing power, graphic cards will continue to advance in technology to meet these demands. The world of graphics, either through movies, games, or regular software applications, is a huge industry that is pushing graphical processing to it's limits. Graphical pipelines are annual being added to newer models of graphic cards to increase processing speeds. Also, with the added potential of processing graphics, displays such as monitors, projectors, or even televisions will be affected by the graphic cards to meet the demand of displaying higher quality images. Graphical technology has increased exponentially since the first computers were invented and will continue to do so meeting the demands for more life-like graphics in applications.

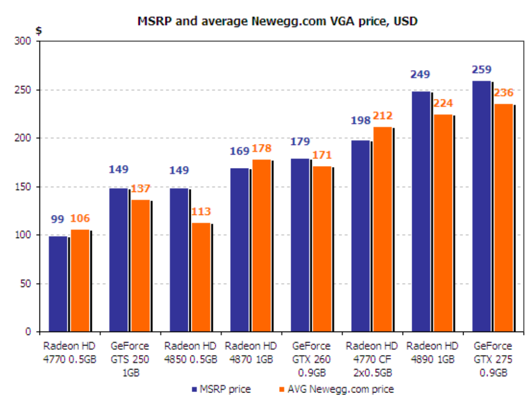



NVIDIA
geFORCE4
VisionTek
4.7


Via K8M800 Northbridge Chipset
DDR DIMM Memory Slots
Socket 754 Connector
ATX Power Connector
Back Panel Connectors
FDD Header
IDE Headers
AGP Slot
Serial ATA Headers
PCI Slots
Via VT8237 Southbridge Chipset
Audio Header
CD-In Header
USB 2.0 Headers



PRICES FOR GRAPHIC CARDS



COMPARISON TABLES FOR AMD AND NVIDIA

AMD ATI Chips						
CHIP	CORE CLOCK	MEMORY CLOCK	MEMORY INTERFACE	MEMORY TRANSFER RATE	PIXELS PER CLOCK	DIRECTX
Radeon HD 5450	650 MHz	800 MHz(DDR2) or 1.6 GHz (DDR3)	64-bit	6.4 GB/s(DDR2) or 12.8 GB/s (DDR3)	80 *****	11
Radeon HD 5550	550 MHz	(varies)	128-bit	(varies)	320 *****	11
Radeon HD 5570	650 MHz	1.8 GHz	128-bit	28.8 GB/s	400 *****	11
Radeon HD 5670	775 MHz	4 GHz	128-bit	64 GB/s	400 *****	11
Radeon HD 5750	705 MHz	4.6 GHz	128-bit	73.6 GB/s	720 *****	11
Radeon HD 5770	850 MHz	4.8 GHz	128-bit	76.8 GB/s	800 *****	11
Radeon HD 5830	800 MHz	4 GHz	256-bit	128 GB/s	1,120 *****	11
Radeon HD 5850	725 MHz	4 GHz	256-bit	128 GB/s	1,440 *****	11
Radeon HD 5870	850 MHz	4.8 GHz	256-bit	153.6 GB/s	1,600 *****	11
Radeon HD 5970	725 MHz	4 GHz	256-bit	128 GB/s	1,600 *****	11

NVIDIA Chips						
CHIP	CORE CLOCK	MEMORY CLOCK	MEMORY INTERFACE	MEMORY TRANSFER RATE	PIXELS PER CLOCK	DIRECTX
GeForce GTX 260 ***	576 MHz / 1,242 MHz	2 GHz	448-bit	112 GB/s	192	10
GeForce GTX 260/216	576 MHz / 1,242 MHz	2 GHz	448-bit	112 GB/s	216	10

***						
GeForce GTX 275 ***	633 MHz / 1,404 MHz	2.268 GHz	448-bit	127 GB/s	240	10
GeForce GTX 280 ***	602 MHz / 1,296 MHz	2.21 GHz	512-bit	141.7 GB/s	240	10
GeForce GTX 285 ***	648 MHz / 1,476 MHz	2.48 GHz	512-bit	159 GB/s	240	10
GeForce GTX 295 ** ***	576 MHz / 1,242 MHz	2 GHz	448-bit	112 GB/s	240	10
GeForce GTX 310 ***	589 MHz / 1,402 MHz	1 GHz	64-bit	8 GB/s	16	10.1
GeForce GTX 465 ***	607 MHz / 1,215 MHz	3,206 MHz	256-bit	102.6 GB/s	352	11
GeForce GTX 470 ***	607 MHz / 1,215 MHz	3,348 MHz	320-bit	133.9 GB/s	448	11
GeForce GTX 480 ***	700 MHz / 1,401 MHz	3,696 MHz	384-bit	177.4 GB/s	480	11