How AGP Works (by Jeff Tyson and Robert Valdes)

http://computer.howstuffworks.com/agp.htm

It all started in 1973, when Xerox completed the Alto, the first computer to use a graphical user interface. This innovation forever changed the way the people work with their computers.

Today, every aspect of computing, from creating animation to simple tasks such as word processing and e-mail, uses lots of graphics to create a more intuitive work environment for the user. The hardware to support these graphics is called a graphics card. The way this card connects to your computer is key in your computer's ability to render graphics. In this article, you will learn about AGP, or Accelerated Graphics Port. AGP enables your computer to have a dedicated way to communicate with the graphics card, enhancing both the look and speed of your computer's graphics.

Get Off the PCI Bus

In 1996, Intel introduced AGP as a more efficient way to deliver the streaming video and real-time-rendered 3-D graphics that were becoming more prevalent in all aspects of computing. Previously, the standard method of delivery was the Peripheral Component Interconnect (PCI) bus. The PCI bus is a path used to deliver information from the graphics card to the central processing unit (CPU). A bus allows multiple packets of information from different sources to travel down one path simultaneously. Information from the graphics card travels through the bus along with any other information that is coming from a device connected to the PCI. When all the information arrives at the CPU, it has to wait in line to get time with the CPU.

This system worked well for many years, but eventually the PCI bus became a little long in the tooth. The Internet and most software were more and more graphically oriented, and the demands of the graphics card needed priority over all other PCI devices.

AGP is based on the design of the PCI bus; but unlike a bus, it provides a dedicated point-to-point connection from the graphics card to the CPU. With a clear path to the CPU and system memory, AGP provides a much faster, more efficient way for your computer to get the information it needs to render complex graphics. In the next section, we'll see how this is done.

AGP Graphics Rendering

AGP is built on the idea of improving the ways that PCI transports data to the CPU. Intel achieved this by addressing all of the areas where PCI transfers were causing data bottlenecks in the system. By clearing the traffic jams of data, AGP increases the speed at which machines can render graphics while using the system's resources more efficiently to reduce overall drag.

Here's how:

- Dedicated Port - There are no other devices connected to the AGP other than the graphics card. With a dedicated path to the CPU, the graphics card can always operate at the maximum capacity of the connection.
- Pipelining - This method of data organization allows the graphics card to receive and respond to multiple packets of data in a single request. Here's a simplified example of this:

  > With AGP, the graphics card can receive a request for all of the information needed to render a particular image and send it out all at once. With PCI, the graphics card would receive information on the height of the image and wait... then the length of the image, and wait... then the width of the image, and wait... combine the data, and then send it out.

- Sideband addressing - Like a letter, all requests and information sent from one part of your computer to the next must have an address containing "To" and "From." The problem with PCI is that this "To" and "From" information is sent with the working data all together in one packet. This is the equivalent of including an address card inside the envelope when you send a letter to a friend: Now the post office has to open the envelope to see the address in order to know where to send it. This takes up the post office's time. In addition, the address card itself takes up room in the envelope, reducing the total amount of stuff you can send to your friend.

  With sideband addressing, the AGP issues eight additional lines on the data packet just for addressing. This puts the address on the outside of the envelope, so to speak, freeing up the total bandwidth of the data path used to transfer information back and forth. In addition, it unclogs system resources that were previously used to open the packet to read the addresses.

PCI Graphics Rendering: Wasting RAM

Speed is not the only area where AGP has bested its predecessor. It also streamlines the process of rendering graphics by using system memory more efficiently.

Any 3-D graphic you see on your computer is built by a texture map. Texture maps are like wrapping paper. Your computer takes a flat, 2-D image and wraps it around a set of parameters dictated by the graphics card to create the appearance of a 3-D image. Think of this as wrapping an invisible box with wrapping paper to show the size of the box. It is important to understand this because the creation and storage of texture maps is the main thing that drains the memory from both the graphics card and the system overall.

With a PCI-based graphics card, every texture map has to be stored twice. First, the texture map is loaded from the hard drive to the system memory (RAM) until it has to be used. Once it is needed, it is pulled from memory and sent to the CPU to be processed. Once processed, it is sent through the PCI bus to the graphics card, where it is stored again in the card's framebuffer. The framebuffer is where the graphics card holds the image in storage once it has been rendered so that it can be refreshed every time it is needed. All of this storing and sending between the system and the card is very draining to the overall performance of the computer.

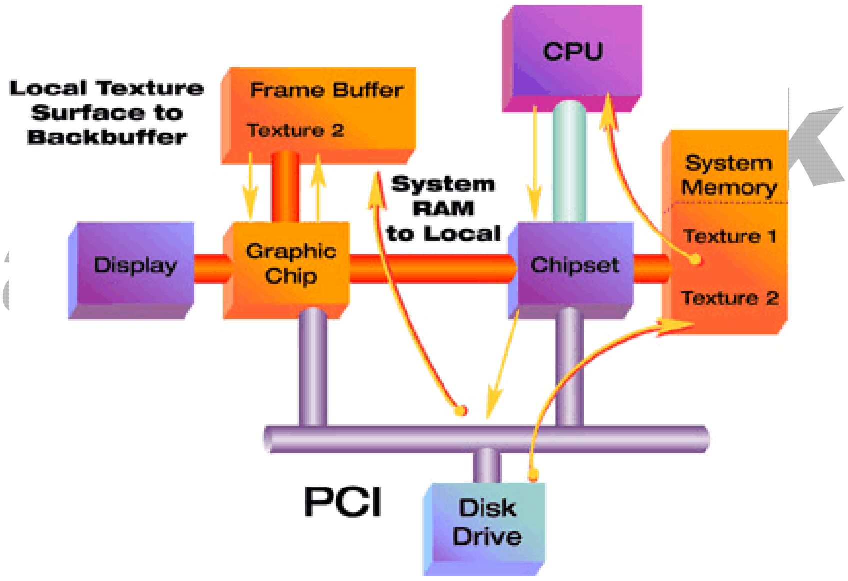

Photo courtesy Intel Corporation
With PCI, texture maps are loaded from the hard drive to system memory, processed by the CPU and then loaded into the framebuffer of the graphics card.

AGP Memory Improvements

AGP improves the process of storing texture maps by allowing the operating system to designate RAM for use by the graphics card on the fly. This type of memory is called AGP memory or non-local video memory. Using the much more abundant and faster RAM used by the operating system to store texture maps reduces the number of maps that have to be stored on the graphics card's memory. In addition, the size of the texture map your computer is capable of processing is no longer limited to the amount of RAM on the graphics card.

The other way AGP saves RAM is by only storing texture maps once. It does this with a little trickery. This trickery takes the form of a chipset called the Graphics Address Remapping Table (GART). GART takes the portion of the system memory that the AGP borrows to store texture maps for the graphics card and re-addresses it. The new address provided by GART makes the CPU think that the texture map is being stored in the card's framebuffer. GART may be putting bits and pieces of the map all over the system RAM; but when the CPU needs it, as far as it's concerned the texture map is right where it should be.

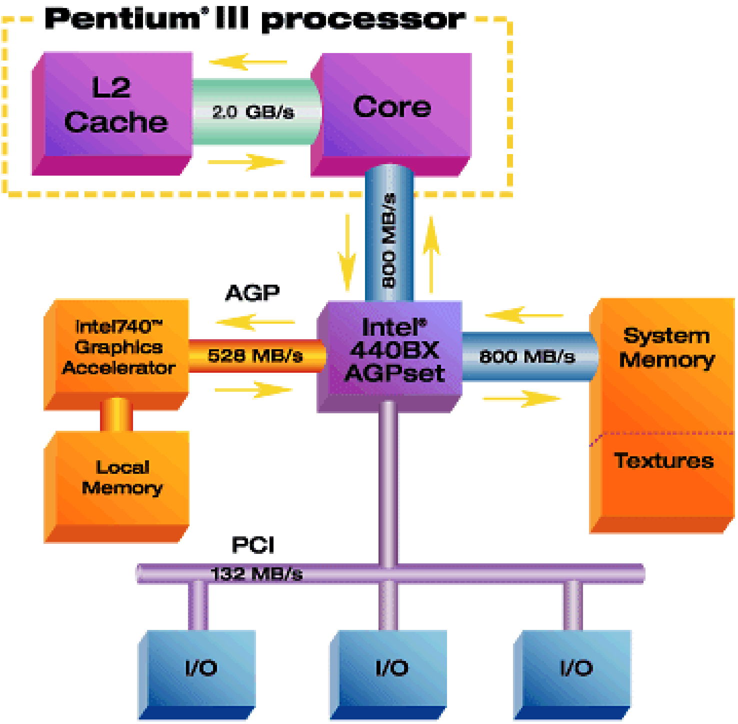

Photo courtesy Intel Corporation

Diagram of the standard architecture of a Pentium III-based system using AGP