Technofile: graphics chips

Whether you're a gamer with a passion for graphics-intensive shoot-'em-ups or a creative designer, a top-notch graphics card is a must. But, says Robin Morris, it's the actual graphics chip that determines whether generated images are more pure than pixellated

Built for gamers but conceived by maths graduates, the graphics card can inspire awe and wonder in its users. Modern graphics cards are so sophisticated that many of them have more onboard transistors than a processor and we could easily devote an entire issue of *PC Advisor* explaining the technology. Instead we're going to take a whistle-stop tour round the graphics chip roadmap, explaining the significance of such terms as transform and lighting, texture mapping as well as core and clock speeds.

Building block of 3D life

When choosing a graphics card it's the actual chip that's the make-or-break factor. Graphical images are constructed from millions and millions of polygons. The speed at which your chip can draw, colour and light all of the polygons determines the overall effect.

A polygon is defined by a number of vertices and each vertex comes with a set of data, defining such attributes as its co-ordinates, weight, colour and texture details. Imagine the vertices as little dots being placed on the screen. Join the dots together and you have a polygon. Generally each polygon consists of three vertices, making it triangular in shape. This is why manufacturers tend to use the terms 'polygon' and 'triangle' interchangeably.

An important part of generating graphics, the transform function makes objects appear to move or change shape – for example, rotating, scaling (making the object bigger or smaller to give a feeling of depth) and translation (altering an object's position).

The term 'lighting' is self-explanatory – applying colour effects to the scene in order to evoke a particular atmosphere, whether it's the flicker of a torch inside a cave or a room bathed in bright light.

Traditionally graphics chips left much of the hard work to the PC processor. nVidia was the first company to release a chip with hardware transform and lighting. The GeForce 256 had the time, resources and specialist knowhow to make an excellent job of generating graphics. And because the processor had a smaller workload, games programmers could find other activities to occupy the PC with – such as creating larger in-depth worlds or more intelligent characters.

ATI swiftly followed with its own version, the charisma engine, which debuted in the Radeon 256. Impressive though the hardware transform and lighting technology was, its 'fixed function' nature cut down on creativity. Once the programmer had defined his vertices and polygons and sent them off to the 3D pipeline, he had little control over how the transform and lighting engine would interpret the information. DirectX 8.0's programmable vertex and pixel shaders remedied this.

These small programs are capable of intercepting and modifying vertex and pixel data. The extra power allows programmers to realistically simulate effects like shadows, reflections, rippling water and natural character animation.

The original specifications were fairly limited in scope, but the 2.0 upgrade (incorporated within version 9.0 of DirectX) has added far more functionality. In the case of pixel shaders, far more texture maps can now be used.

Although ATI was the first manufacturer to provide version 2.0 vertex and pixel shaders, nVidia has gone far beyond the Microsoft remit and prefers to designate its glorified specifications as version 2.0+. Enticing though this may sound, it does prompt an important question – will any of this make any difference to today's gaming experience? Expect it to be a good 12-18 months before we see any games getting anywhere near to using the 2.0 specifications, never mind 2.0+.

APIs explained

The API (application programming interface) was originally a means of coping with the ever-changing PC technology. Whereas a games console can't be upgraded and can, therefore, be relied upon to have the same graphics controller for its entire lifetime, a PC is in a constant state of flux. Every time a new graphics chip comes out, games code may need to be tweaked and rewritten to ensure compatibility.

APIs provide a library of readymade functions and tools that programmers can select and use in creating their games. And provided that the games are written to certain guidelines, they should work with any graphics card compatible with that API.

APIs like OpenGL and Microsoft's constantly evolving DirectX standard are so important chip manufacturers often design their new products around the features set of upcoming versions. This has limitations, though.

Although Microsoft rushes out new releases of DirectX on an annual basis, it's still a struggle to keep up with the latest advances in 3D graphics. The programmer also has little say over exactly how API effects will be produced or which parts of the graphics card will be used. instead, they have to rely on either the API or the implementation of the graphics card drivers.

And since an API consists of little more than a standard-issue programmers' toolkit, it's hard to make your game stand out from the crowd when everybody's using exactly the same tools. Luckily, with version 8.0 of its DirectX API, Microsoft lets games programmers exercise more creativity by equipping them with 'programmable' vertex and pixel shaders. These were quickly implemented by nVidia in its nFiniteFX engine and ATI's Smartshader.

Shady character

However many polygons you use in an object, you won't be able to breathe life into the image without shading and texturing. Shading amounts to little more than colouring in the polygons. There are more advanced techniques than the rather basic process of 'flat shading' – that is, applying a single colour to each polygon.

In Gouraud shading, a colour is applied to each vertex and a complex algorithm is used to blend the three colours together into one realistic covering. Phong shading works with pixels rather than vertices and, while it's very effective, it is resource-intensive.

Textures (or texture maps) are small pictures that can be stretched across an object to simulate an effect. For example, a cube shaped object can be turned into a brick wall by fitting it with a picture of brickwork. A texture can be as large as you want, but ideally you want it to be roughly the same size as the object it's being applied to. All pixels used to form a texture are referred to as texels.

The problem games designers face is that a texture can be viewed from many angles and distances. Should an object suddenly double in size, the original texture would either have to be stretched to the point where the texture picture was severely distorted or it would only fill a quarter of the object.

Mip mapping is a partial solution and involves making several copies of a texture at different sizes – if a 64x64 wall doubles in size, we can replace the original 64x64 texture with the 128x128 version. Textures take up memory, though, so mip mapping isn't that practical. Instead, programmers use sampling or filtering techniques to 'redraw' the texture on to a larger (or smaller) area.

In the above example, we could split the texture image into a 64x64 grid of coloured squares. By placing a second 64x64 grid over the new object (increasing the size of the squares so that the grid fits perfectly) and transferring the texture one square at a time, we can create a new resized texture.

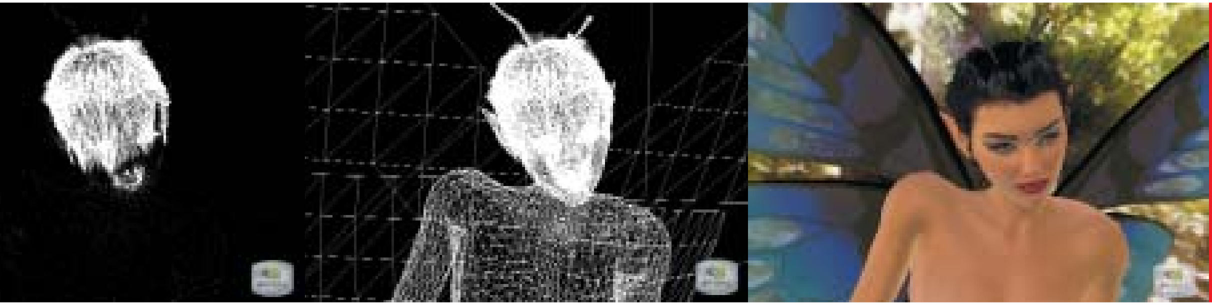
Keep the detail

Unfortunately, this process (called 'point sampling') of turning blocks of one colour into even larger blocks of the same shade means that as the texture size increases the image gets blockier. Previously beautifully defined game objects can break down into a mess of coloured dots.

Bilinear filtering is more realistic. Rather than transfer the grid one square at a time, the PC takes the colours of the four surrounding squares and blends them together to create one hue. Trilinear filtering uses the four-square technique too, but also employs mip mapping to create two different sized textures.

Whereas these other forms of filtering use square-shaped samples, anisotropic filtering takes samples that are rectangular, trapezoidal or parallelogram shaped. By sampling from a variety of different angles, anisotropic filtering produces the best results of all, but it can eat into memory resources.

Another technique, FSAA (full scene anti-aliasing), smooths out the jagged


Building up a graphics image: start with the vertices; join the dots to form polygons; finally add light, shading and textures

Understand the lingo

ATI and nVidia use lots of different words to describe the same technology. We list some of the common phrases and describe what they do.

nVidia technology	ATI technology	Purpose
Transform and lighting engine	Charisma Engine	Fixed function hardware transform and lighting engine
nFiniteFX	Smartshader	Programmable vertex and pixel shaders
NSR (nVidia shading rasterizer)	Pixel tapestry	3D graphics features such as bump mapping
Lightspeed memory architecture	HyperZ	Memory management and compression
Detonator drivers	Catalyst	Driver and multimedia centre
nView	Hydravision	Multiple monitors

edges on graphics images. The FX 5800 Ultra and Radeon 9700/9800 have advanced FSAA modes that can run with little drop in frame rates.

To the core

No matter how many polygons, textures and vertex and pixel shader effects your graphics card can generate, if your games run like a snail in quicksand your pulse won't race when you play them. We want attractive graphics, but we don't want good looks to be at the expense of frame rates.

Graphics cards have two performance accelerators: the core clock speed and the memory clock speed. Most graphics chips come in either a standard or enhanced version, the latter with higher core and memory clock speeds. The enhanced nVidia and ATI cards are marked Ultra and Pro respectively.

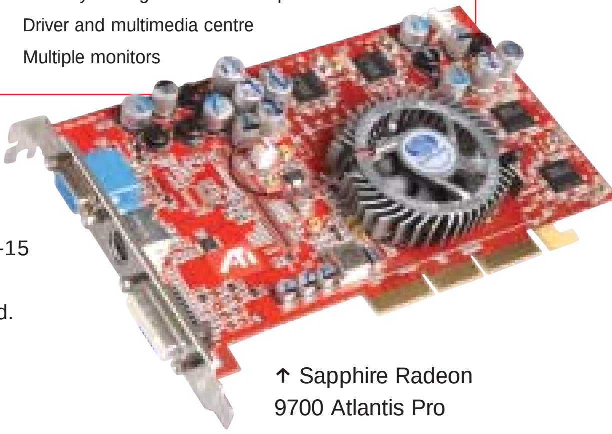
The core clock (sometimes referred to as the GPU clock) determines how quickly the graphics chip is running and is a good indication of performance. The higher the clock speed, however, the more heat is generated. Combine this with the complexity of modern graphics chips and you have a reason for the huge, and sometimes noisy, cooling systems. Individual components are also much smaller on today's chips – around 0.13-15 microns. Less heat is generated and therefore the core speeds can be raised.

Pixels at dawn

The core clock setting also influences the speed of the pixel pipes – the series of processes and components through which graphics data is fed and processed. Modern chips tend to be equipped with more than one pixel pipe, allowing them to process lots of data simultaneously. Each pipe will have at least one texture unit and the more of these you have the more textures you can process per clock cycle.

So although the Radeon 9700 chip has eight pixel pipes as opposed to the GeForce4's four pipes, the latter chip makes up some ground by having twice as many texture units per pipe. Ultimately, the 9700 is faster, partly due to its superior core speed. Unfortunately, a chip like the GeForce FX 5200 loses out on all counts and its four pixel pipes and one shader unit are inadequate for today's needs.

→ Command & Conquer: those Generals look so much better with a cutting-edge graphics card



↑ Sapphire Radeon 9700 Atlantis Pro

Version 9.0 of the DirectX API introduced 'floating point' pixel pipes. Previously, graphics chips used integer pipes that, faced with a number like 8.4333372, would have no alternative but to round it up or down to a whole number. However, the new floating point pixel pipes can deal freely with decimal places.

This may seem trivial, but floating point calculations are far more complex and result in more realistic graphics. The larger range of numerical values also ushers in the age of 128bit colour, meaning that future games titles will have access to a greater selection of colours and shades when creating graphics images.

Manufacturers often boast about high fill rates, supposedly a measure of the amount of graphics a chip can draw per second. The pixel fill rate is the most common and can be calculated by multiplying the core clock speed by the number of pixel pipes.

The result will be in megapixels, so you'll need to divide by 1,000 to convert it to gigapixels. In the case of the Radeon 9800 Pro, the fill rate would amount to 3.04 gigapixels per second.

Multiplying this figure by the number of texture units per pipe will give you the texel fill rate, which determines how many textures can be generated. In the case of

Features comparison: graphics chips

Graphics chip	Approx price (ex VAT)	Transistors/ process	Memory bus/ configuration	No of pixel pipes	Core clock speed	Memory clock speed	Memory bandwidth	Pixel fill rate (gigapixels per sec)
ATI Radeon 9800 Pro	£300-£350	107 million/ 0.15 micron	256bit/ 128MB DDR	8 (1 texturing unit per pipe)	380MHz	340MHz (680MHz)	21.8GBps	3.04
nVidia GeForce FX 5800 Ultra	£275-£375	125 million/ 0.13 micron	128bit/ 128MB DDR II	8 (1 texturing unit per pipe)	500MHz	500MHz (1,000MHz)	16GBps	4
ATI Radeon 9700 Pro	£190-£245	107 million/ 0.15 micron	256bit/ 128MB DDR	8 (1 texturing unit per pipe)	325MHz	310MHz (620MHz)	19.8GBps	2.6
nVidia GeForce4 Ti 4800	£130-£200	63 million/ 0.15 micron	128bit/ 128MB DDR	4 (2 texturing units per pipe)	300MHz	325MHz (650MHz)	10.4GBps	1.2
nVidia GeForce4 Ti 4200 8x	£100-£150	63 million/ 0.15 micron	128bit/ 128MB DDR	4 (2 texturing units per pipe)	250MHz	250MHz (500MHz)	8GBps	1
ATI Radeon 9600 Pro	£100-£120	75 million/ 0.13 micron	128bit/ 128MB DDR	4 (1 texturing unit per pipe)	400MHz	300MHz (600MHz)	9.6GBps	1.6
nVidia GeForce FX 5200 Ultra	£70-£80	45 million/ 0.15 micron	128bit/ 128MB DDR	4 (1 texturing unit per pipe)	325MHz	325MHz (650MHz)	10.4GBps	1.3
ATI Radeon 9200 Pro	£50-£65	not specified/ 0.15 micron	128bit/ 128MB DDR	4 (1 texturing unit per pipe)	250MHz	200MHz (400MHz)	6.4GBps	1
nVidia GeForce3 Ti 200	£40-£50	57 million/ 0.15 micron	128bit/ 64MB DDR	4 (2 texturing units per pipe)	175MHz	200MHz (400MHz)	6.4GBps	0.7

the Radeon 9800 Pro only one texture unit is assigned to each pipe, so the pixel and texel fill rates would be identical. However, the GeForce4 Ti 4800 comes with four pixel pipes, each one equipped with two texture units.

Remember that these are peak figures and assume virtually no levels of detail and low resolutions. Even with low-quality graphics, fill rates are often impossible to attain simply because the chip lacks the memory bandwidth necessary to process so much information.

Bandwidth bother

The memory clock refers to RAM speed. Most RAM is now the DDR variety which can perform twice as many actions per clock cycle as older SDR RAM. If the memory clock is 300MHz, DDR RAM will double this to 600MHz. Both numbers are often given, with the larger figure included in brackets. DDR II is similar but runs at a higher raw memory clock speed.

With all the extra duties being taken on by GPUs, you might need to relieve the pressure by increasing memory bandwidth. Measured in GBps (gigabytes per second), memory bandwidth amounts to breathing space. The more detail you have (higher resolutions, larger numbers of textures, 32bit colour and upwards), the more memory bandwidth you'll need. This, rather than the fill rate, is likely to prove the chip's greatest liability.

To calculate the memory bandwidth, multiply the memory clock speed (double the clock if DDR RAM is used) by the size of the memory bus. Now, to convert the memory bus from bits to bytes, divide it

Put to the test: 3D graphics-intensive games

Graphics chip	Overall position	Unreal Tournament 2003 (resolution)			Quake III (resolution)		
		1,024x768	1,280x1,024	1,600x1,200	1,024x768	1,280x1,024	1,600x1,200
ATI Radeon 9800 Pro	1st	159.7fps	147.7fps	123.4fps	213.2fps	210.4fps	194.7fps
nVidia GeForce FX 5800 Ultra	2nd	154.2fps	143.9fps	122.3fps	215.9fps	212.2fps	195.2fps
ATI Radeon 9700 Pro	3rd	158.1fps	138fps	106.5fps	213fps	208.9fps	189.3fps
nVidia GeForce4 Ti 4800	4th	148.2fps	109.6fps	80.4fps	208.6fps	191fps	165.3fps
nVidia GeForce4 Ti 4200 8x	5th	133.7fps	92.1fps	66.7fps	203.2fps	169.9fps	131.4fps
ATI Radeon 9600 Pro	6th	130fps	90.1fps	61.5fps	201fps	164.8fps	122.6fps
nVidia GeForce FX 5200 Ultra	7th	70.6fps	45.7fps	33fps	183.5fps	123.2fps	88fps
ATI Radeon 9200 Pro	8th	62.8fps	41.5fps	30.4fps	176.1fps	116.4fps	69.2fps
nVidia GeForce3 Ti 200	9th	36.3fps	32.9fps	20fps	185.8fps	128.1fps	99.9fps

fps = frames per second

	Texel fill rate (gigatexels per sec)	DirectX version	AGP version	Example card & website	Chip review
	3.04	9.0	8x	Sapphire Technology (www.sapphiretech.com)	Fastest overall card but also extremely expensive. For dedicated gamers only
	4	9.0	8x	Gainward (www.gainward.com), MSI (www.msi.com.tw)	Good overall speed. This card's failings are a high price tag and exorbitant levels of operating noise
	2.6	9.0	8x	Gigabyte (http://uk.giga-byte.com), Sapphire Technology	Only a fraction down on the other P800 Pro and 5800 for speed, but falling price makes it a fantastic buy
	2.4	8.1	8x	PNY (www.pny.co.uk)	The 8x version of the old Ti 4600 offers good performance and value for money
	2	8.1	8x	Abit (www.abit.com.tw), AOpen (www.aopen.nl)	Extremely cheap considering the high performance levels. Great if you don't play games at maximum detail levels
	1.6	9.0	8x	Hercules (www.hercules.com)	Value for money, but meagre pixel pipe and texture unit configuration leaves it trailing the Ti 4200 8x
	1.3	9.0	8x	MSI, Asus (www.asus.com.tw)	Solid card with modest configurations. Adequate but you're better off paying the extra for a Ti 4200
	1	8.1	8x	Sapphire Technology	A few driver issues in the pre-release version. Slightly too cheap for comfort
	1.4	8.1	4x	Gainward	Low price but 64MB of memory leaves it struggling

by eight – so for 128bit and 256bit buses you would use 16 bytes and 32 bytes respectively. Thus the Radeon 9800 Pro has a memory bandwidth of 21.8GBps.

Wheels on the bus

RAM chips have their own speed ratings (measured in nanoseconds), which refer to the memory clock speed at which you can run the RAM. For example, 4ns memory chips can run at up to 500MHz. With a 128bit memory bus, this would produce a bandwidth of 8GBps.

By replacing 4ns RAM with 3.6ns or 3.3ns chips (supporting 556MHz and 600MHz respectively), memory bandwidth could be stretched to 8.9 or 9.6GBps. If memory bandwidth exceeds the official graphics chip specifications, it could be that the card maker is using faster RAM.

The raw figure doesn't tell the whole story, though. Some manufacturers have made the memory bus more efficient by breaking it down into smaller chunks – the GeForce3 uses a 128bit bus, for instance, but increases performance by splitting it into four individual 32bit buses.

Luck of the draw

As well as X and Y co-ordinates, graphics are assigned a Z co-ordinate. This value tells the graphics card how close the polygons are, and whether they are hidden from the view of the player – for example, a room hidden behind a door.

← You're much more likely to shoot straight playing a game such as No One Lives Forever 2, shown here, when you have crisp detail provided by the right graphics acceleration

If you can't see the polygon there's no point in wasting resources drawing it, so the culling process removes any that won't be visible. This 'overdraw' works via such technologies as nVidia's Z-Occlusion and ATI's HyperZ. Texture compression also increases bandwidth.

Buying a motherboard with an 8x AGP port increases the bandwidth from 1GBps to 2.1GBps. In testing, however, the difference is rarely more than 2fps (frames per second). Most cards are 8x and an 8x AGP model should work in a motherboard with a 4x AGP socket, so this won't really be a factor. If in doubt, simply specify 8x.

As a rule, higher resolutions and detail levels require more memory bandwidth. But while calculating fill rates and memory bandwidth can be useful for comparing graphics chips, it's 'hidden' performance that determines overall speed.

For instance, nVidia claims the strong memory optimisation of its FX 5800 Ultra chip results in an effective doubling of the raw memory bandwidth. In games tests, the 5800 Ultra loses about 3-5fps to the Radeon 9800 Pro at low resolutions, but catches up at 1,600x1,200 and beyond.

Since the 5800 Ultra has less memory bandwidth but a greater fill rate, the realworld results are almost the opposite of what you would expect. All this goes to prove that, no matter what technology is onboard, and what theoretical figures are generated, the frame rates you can get in today's games are as reliable a test of merit as any. ■

Unsure of a technical term? Find out exactly what it means in our searchable Glossary which is on the cover disc