



Radeon® X1900

Technology Brief



Radeon X1900 Technology Brief

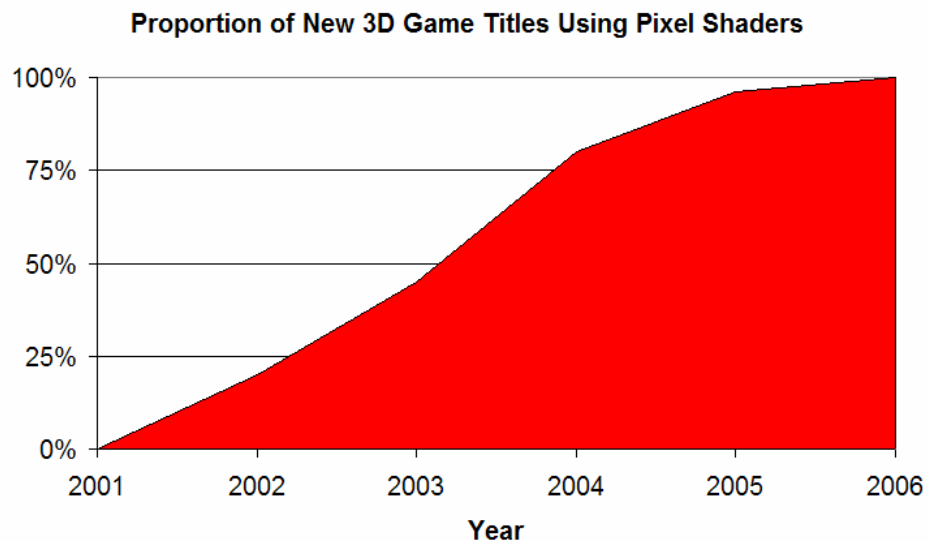
The Radeon X1900 series sets the new standard for high-end gaming GPU performance. It is the first series of PC graphics products to feature 48 pixel shader processors, twice as many as any previous technology. This huge leap in pixel shading ability is empowering developers to use new features and techniques that have been precluded by performance limitations until now, while still using the industry standard DirectX® 9 API and Shader Model 3.0.

The maturation of 90nm process technology allows the Radeon X1900 to take advantage of unprecedented transistor density. With over 380 million transistors, the design is able to combine all of the key innovations of the Radeon X1000 series (including Ultra-Threading, HDR with Anti-Aliasing, image quality improvements, Avivo video & display technologies, and more) with massively increased shader processing power and other interesting new features like Fetch4 shadow map acceleration.

This paper describes some of the research behind the new design, as well as the new capabilities of the Radeon X1900 and how they will drive the next generation of 3D gaming technology.

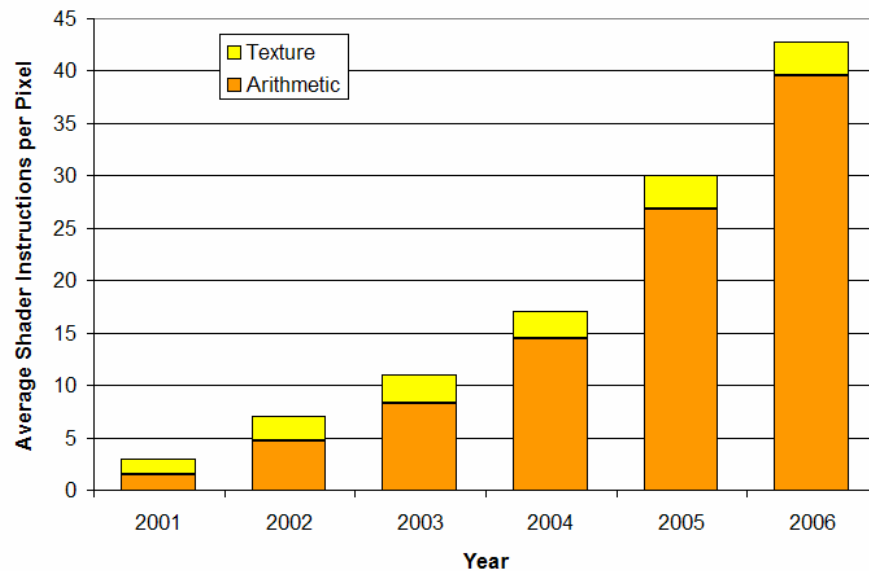
Shader Trends

Pixel shading has become the key performance bottleneck in today's latest game engines. This has been the result of a steady increase in the number of shader instructions executed per rendered pixel over the past few years. Since the introduction of programmable shaders in 2001 with the release of the DirectX 8 API from Microsoft®, shaders have not only become ubiquitous in games, but their complexity has grown by an average factor of about 1.8x per year.



Pixel Shader Adoption in PC Game Titles

A more detailed analysis of the shaders in 3D games reveals that not only have they increased in size, but their composition has been changing as well. Shader instructions can be divided into two general classes: texture operations, which fetch data from memory, and arithmetic operations, which perform mathematical manipulation of data. While early shaders were divided roughly equally between these two types of instructions, more recent shaders tend to have a much higher proportion of arithmetic instructions. In the latest games, the average ratio of arithmetic to texture operations is approaching 5:1, and is projected to continue increasing in the next generation of game titles in 2006 and beyond.



Pixel Shader Instruction Counts in PC Game Titles

Another observation is that in recent games, a majority of the pixels processed use bilinear filtering or point sampling from integer textures, or no texture lookups at all. These shader operations can be executed by each texture unit in one clock cycle or less. This balances against the remainder of the pixels which require trilinear filtering, anisotropic filtering, and/or floating point texture lookups, which require more than one clock cycle to execute.

One important difference between arithmetic and texture operations is that the latter are heavily dependent on external factors for performance, including graphics memory size and bandwidth. Adding more texture units can become fruitless without commensurate increases in these resources. Memory and bandwidth are in general more costly to scale than arithmetic operations, which depend only on the number of processing units that can be fit into the GPU.

Procedural texturing is one important technique that can take advantage of this trend. Pixel shader programs can be used to generate textures mathematically based on a set of artist-controlled input parameters. This has the potential to dramatically reduce the amount of graphics memory and bandwidth required to store texture data. Alternatively, shader programs could be used to add variation and detail to existing textures, thus reducing the number of different texture maps that need to be stored in memory.

Increasing shader processing power also opens up greater opportunities for the GPU to share CPU workload when it becomes a bottleneck. Physical simulations of things like particle systems, cloth, fluid flow, and animation can be mapped efficiently to GPUs in many cases using shaders. However, this can steal precious resources away from standard graphics rendering. The more shader processors are available to a system, the more likely it becomes that load balancing can be used to increase overall frame rates.

Shader Model 3.0 Done Right

The Shader Model 3.0 specification, part of the DirectX 9 API, includes capabilities that broaden the range of possibilities available to game developers, and offer new opportunities for performance optimization. A new generation of graphical techniques and effects that may have been possible but not practical to implement in earlier shader models (due to unacceptable performance or complexity issues) can now be explored – as long as the available graphics hardware is designed to handle Shader Model 3.0 efficiently.

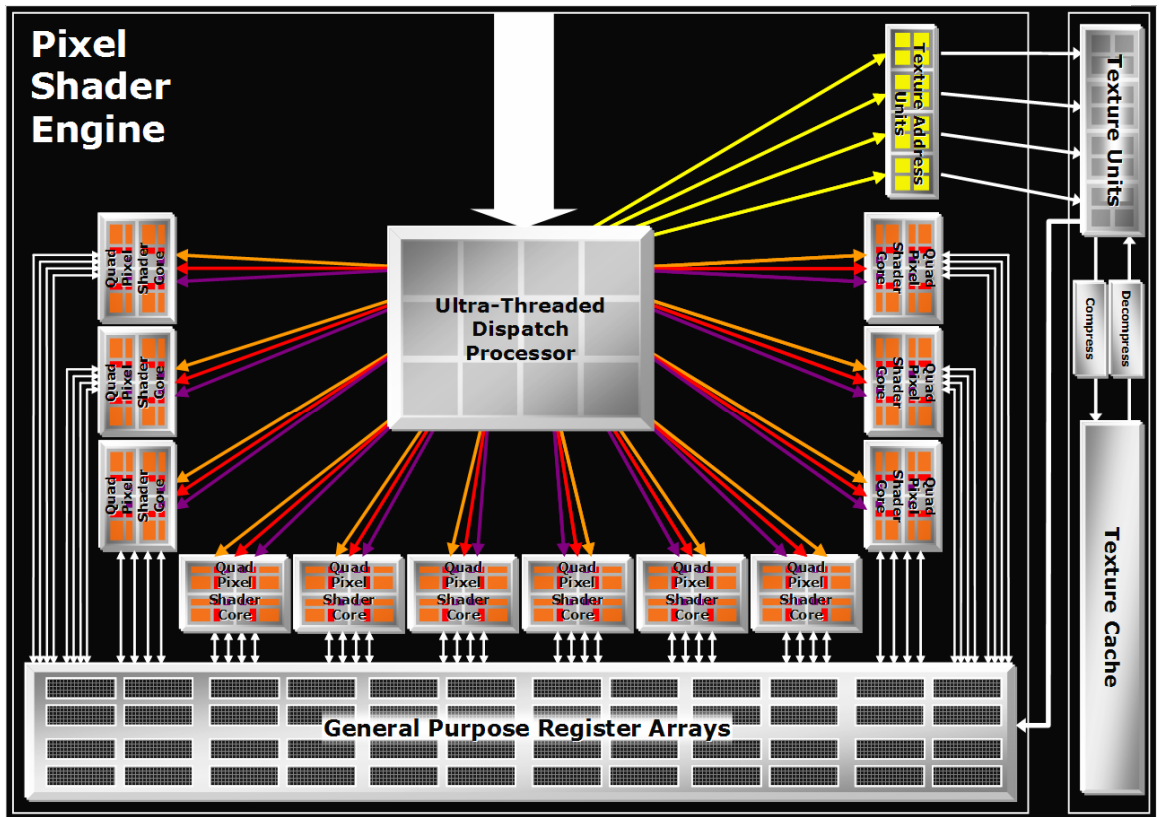
The most important new capability of Shader Model 3.0 is dynamic flow control for pixel shaders. This allows a pixel shader program to execute different code paths, or loop over portions of code multiple times, according to conditions determined for each individual pixel. In earlier shader models, all of the instructions and texture fetches in a shader had to be executed once for each pixel, whether they were required or not. Flow control allows much more sophisticated effects to become practical by executing them only on the specific pixels that require them, while using simpler effects elsewhere.

Getting dynamic pixel shader flow control right in hardware can be tricky, since GPUs in general owe their high speed processing capabilities to extensive parallelism. In other words, they are built to execute a series of operations on multiple pieces of data at once. Dynamic flow control results in different pieces of data executing different operations, which can counteract the benefits of parallelism.

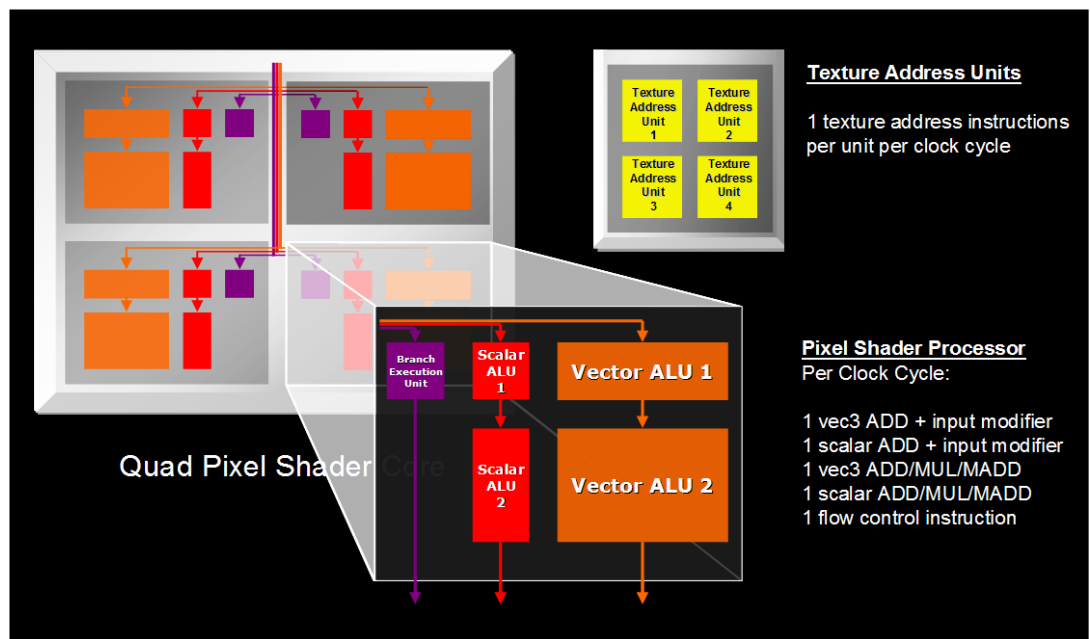
The Radeon X1000 family of GPUs feature Ultra-Threading technology to achieve an optimal balance between fast dynamic flow control and extensive parallelism. Through a combination of large thread counts, small thread sizes, dedicated branch execution units, and a large, high performance register array, this technology achieves pixel shading performance many times faster than competing products. This unlocks a wide variety of new visual effects that would not have been practical to use before.

Radeon X1900 Pixel Shader Architecture

Today's GPU designs include a number of specialized processing units to perform the different types of shader operations. In the Radeon GPU family, arithmetic operations are handled by pixel shader processors consisting of a set of ALUs (Arithmetic Logic Units), while texture operations are handled by dedicated texture units. To ensure optimal use of processing resources, the number of each type of unit present in the GPU should match as closely as possible the expected ratio of instructions it expects to handle.



Radeon X1900 Pixel Shader Engine



Radeon X1900 Pixel Shader Processor Detail

Each pixel shader processor in the Radeon X1900 can handle anywhere from 1 to 5 shader instructions per clock cycle in its various ALUs. Dedicated branch execution units are included to reduce the performance overhead of flow control instructions. Each texture unit and texture address unit can process up to 4 texture fetches per clock cycle.

These units are assigned tasks by the Ultra-Threaded Dispatch Processor, which is constantly seeking opportunities to re-order instructions to achieve maximum utilization of these ALUs. It also makes use of a large number of simultaneous threads to hide texture fetch latency, which can occur when attempting to access data that is not immediately available in the texture cache. Thread sizes are kept small to maximize the benefits of branching operations.

The key observation from the shader analysis described earlier is that to design a GPU that will handle both current and future games as efficiently as possible, it is most important to focus on improving the processing speed of arithmetic pixel shader operations. The Radeon X1900 places a heavy emphasis here, with 48 pixel shader processors providing three times the arithmetic processing power of previous flagship GPUs. By adding 20% more transistors, shader processing power is increased by 200%. The 3:1 ratio of arithmetic to texture units provides the ideal balance for current and future 3D performance.

Shadow Map Acceleration and Fetch4

Texture lookups have long been a common operation in 3D rendering. One widely used class of techniques that is placing a heavier weight on the importance of texture filtering is shadow mapping. This method of rendering shadows works by first rendering the scene from the point of view of a shadow-casting light source. The results are not displayed, but instead stored in a special shadow map texture where each value represents the distance of the nearest object to the light source. The scene is then rendered from the standard viewpoint, and each pixel is checked against the shadow map to determine if there are any objects between it and the light source. If the result is true, the pixel is in shadow and can be darkened, otherwise it is lit normally.

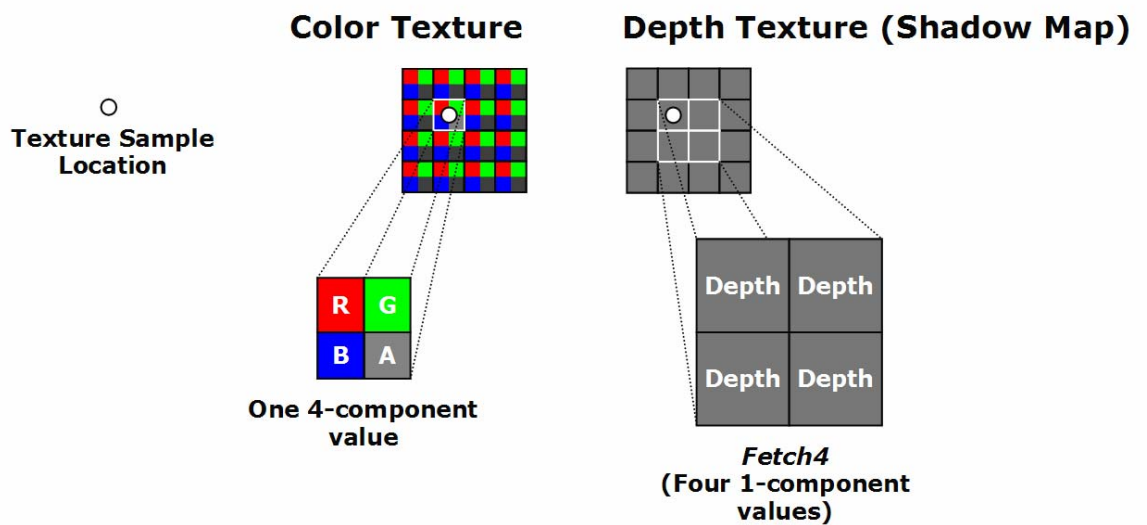
One limitation of shadow maps is that they normally create hard-edged shadows. In the real world, shadows tend to have softer edges. Techniques that create soft-edged shadows often work by filtering the shadow map. This can be done by taking a number of samples and then combining them in a pixel shader. Using a larger number of samples can result in higher quality shadows, but also requires a large number of texture lookups, which can hurt performance.

Dynamic branching can be used to improve the performance of this shadow rendering technique by detecting pixels that lie on or near shadow edges. These pixels can then use a high quality filter with many texture samples, while other pixels use just a single texture lookup to determine if they are in or out of shadow.



ATI Parthenon demo, including soft shadows with filtered shadow maps

To further facilitate this technique, the Radeon X1900 includes a new texture sampling feature known as Fetch4. It works by exploiting the fact that most textures are composed of color values, each consisting of four components (Red, Green, Blue, and Alpha or transparency). The texture units are designed to sample and filter all four components from one texture address simultaneously. However, when looking up different types of textures with single-component values (such as shadow maps), Fetch4 instead allows four values from adjacent addresses to be sampled simultaneously. This effectively increases the texture sampling rate by a factor of 4.



With Ultra-Threading technology providing fast flow control and Fetch4 providing fast texture lookups, the Radeon X1900 can render attractive soft shadows at speeds approaching those of traditional hard-edged shadow mapping techniques.

High Resolution Gaming

New high definition digital displays with 2 megapixels or more are rapidly becoming more common and affordable. Running the latest games at resolutions such as 1920x1200 (WUXGA), 2048x1536 (QXGA) or even 2560x1600 (WQXGA) places heavy demands on pixel shading, fill rate, and memory bandwidth.

All Radeon GPUs support a Hierarchical Z feature, which is designed to significantly reduce these requirements. It works by detecting and eliminating pixels that will be hidden in the final rendered image, and discarding them before any further processing takes place. However, it requires high speed on-chip memory to function, and this memory is of limited size. Rendering at resolutions higher than this memory was designed to support can reduce the effectiveness of Hierarchical Z.

The Radeon X1900 incorporates 50% more on-chip memory for Hierarchical Z than the Radeon X1800. This ensures that its performance does not drop off precipitously at very high resolutions. Combined with support for 512MB memory configurations and dual integrated dual-link DVI transmitters in the Avivo display engine, the Radeon X1900 provides the best available solution for high resolution gaming.

Previewing Future Performance

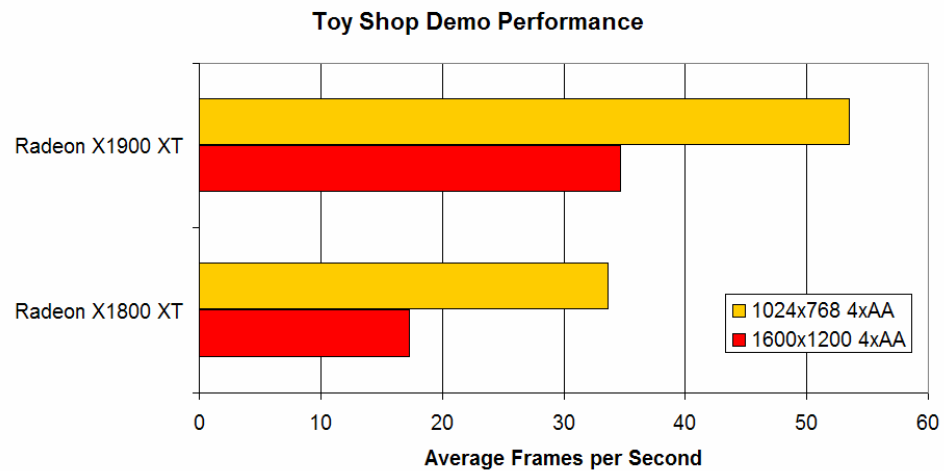
Tests have shown that the Radeon X1900 is the undisputed performance leader in current games that use Shader Model 3.0. However, these games were developed using GPUs with just a fraction of the pixel shading power that is now available. Now that this new level of power is here, what will the future bring?

The Toy Shop demo, available from www.ati.com, provides a compelling example of what is possible with the latest GPUs. It uses a wide range of shader effects that make full use of new capabilities including dynamic pixel shader flow control, HDR lighting, and 3Dc+ texture compression. Some of these effects include parallax occlusion mapping for detailed 3D surfaces, physical simulation of rain and water on the GPU, soft shadows, and much more.



ATI Toy Shop Demo

While this demo runs smoothly on the previous generation Radeon X1800 GPU, the new Radeon X1900 is able to run it at approximately two times the frame rate. This makes it possible to run at higher resolutions and with more sophisticated anti-aliasing methods, delivering incredible image quality.

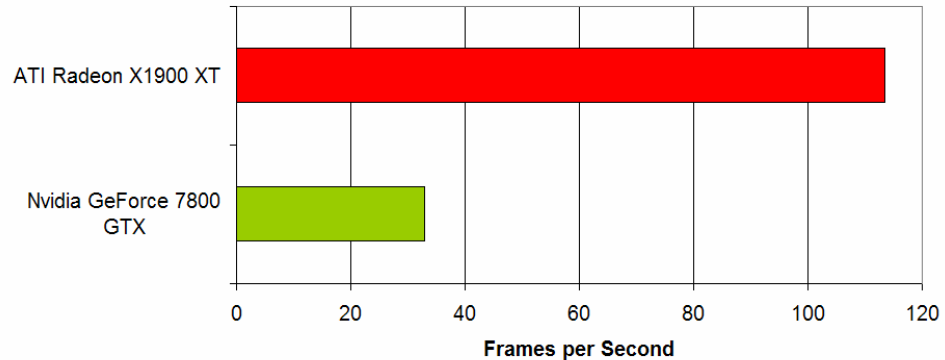


Parallax occlusion mapping is a good example of a technique that is particularly well suited to GPUs like the Radeon X1900, and will see widespread use in the next generation of games. It combines normal mapping with a simple form of ray-tracing to add convincing 3D detail to flat surfaces, complete with realistic reflections and self-shadowing features. It is an iterative process that can take advantage of massive arithmetic processing power and dynamic flow control to greatly enhance the realism of a 3D scene.

One limitation of other techniques like this is that they tend to break down when viewed up close or at sharp angles, resulting in artifacts that ruin the effect. This problem can be solved by increasing the number of iterations per pixel, but this has a major impact on frame rate. Parallax occlusion mapping uses pixel shader flow control to dynamically vary the number of iterations according to the viewing angle and distance, so that only pixels that would otherwise suffer from artifacts get additional iterations.

In the Toy Shop demo, parallax occlusion mapping is used on many of the surfaces in the scene, with up to 50 iterations per pixel while maintaining over 30 frames per second at 1600x1200 resolution with 4x anti-aliasing. This level of performance requires efficient dynamic flow and generous amounts of shader processing power. With Ultra-Threading technology and 48 pixel shader processors, the Radeon X1900 is able to render this effect at frame rates 3-4 times higher than the fastest competing products.

Parallax Occlusion Mapping Performance



DirectX 9.0c SDK Parallax Occlusion Mapping Sample
12-50 samples per pixel with self-shadowing enabled



Parallax Occlusion Mapping

While the level of shading power provided by the Radeon X1900 is new to PC users, it is not completely new to game developers. The ATI-designed GPU in Microsoft's Xbox 360 game console also features 48 shader processors and 16 texture units. The shader processors differ from those in the Radeon X1900 in a number of ways, primarily in that they are unified, meaning there is no distinction between pixel and vertex shader processors. However, there are enough similarities to expect that many effects developed for one platform should be straightforward to port to the other. The important point is that equivalent levels of shader processing power are now available on both the PC and game console platforms, which will help drive graphical innovation on both.



Summary

Based on years of research into the use of pixel shaders in 3D applications, the Radeon X1900 leverages and extends the key features of the Radeon X1000 series. As the first PC GPU with 48 pixel shader processors, its balanced design makes the best use of its transistors to accelerate current games faster than ever before, while simultaneously opening up a huge range of new effects for next-generation titles. New Fetch4 shadow map acceleration and improved Hierarchical Z for high resolution performance round out its new capabilities.

Copyright 2005, ATI Technologies Inc. All rights reserved. ATI and ATI product and product feature names are trademarks and/or registered trademarks of ATI Technologies Inc. All other company and product names are trademarks and/or registered trademarks of their respective owners. Features, availability and specifications are subject to change without notice.

