Technical Brief

NVIDIA GeForce® GTX 200 GPU Architectural Overview

Second-Generation Unified GPU Architecture for Visual Computing

Table of Contents

Figures

Tables

Introduction

In this technical brief we introduce NVIDIA's new GeForce® GTX 200 GPU family, the first GPUs to implement NVIDIA's second-generation unified graphics and computing architecture. The high-end, enthusiast-class GeForce GTX 280 GPU and performance-oriented GeForce GTX 260 GPU are the first members of the GeForce GTX 200 GPU family and deliver the ultimate visual computing and extreme high-definition (HD) gaming experience.

We'll begin by describing architectural design goals and key features, and then dive into the technical implementation of the GeForce GTX 200 GPUs. We assume you have a basic understanding of first-generation NVIDIA unified GPU architecture, including unified shader design, scalar processing cores, decoupled texture and math units, and other architectural features. If you are not well versed in NVIDIA unified GPU architecture, we suggest you first read the Technical Brief titled *NVIDIA GeForce 8800 GPU Architecture Overview*. You can also refer to *Appendix A* for a historical retrospective.

GeForce GTX 200 Architectural Design Goals and Key Capabilities

GeForce GTX 200 GPUs are massively multithreaded, many-core, visual computing processors that incorporate both a second-generation unified graphics architecture and an enhanced high-performance, parallel-computing architecture.

Two overarching themes drove GeForce GTX 200 architectural design and are represented by two key phrases: **"Beyond Gaming"** and **"Gaming Beyond."**

Beyond Gaming means the GPU has evolved beyond being used primarily for 3D games and driving standard PC display capabilities. More and more, GPUs are accelerating non-gaming, computationally-intensive applications for both professionals and consumers.

Gaming Beyond means that the GeForce GTX 200 GPUs enable amazing new gaming effects and dynamic realism, delivering much higher levels of scene and character detail, more natural character motion, and very accurate and convincing physics effects.

The GeForce GTX 200 GPUs are designed to be fully compliant with Microsoft DirectX 10 and Open GL 2.1.

Architectural Design Goals

NVIDIA engineers specified the following design goals for the GeForce GTX 200 GPUs:

- Design a processor with up to twice the performance of GeForce 8800 GTX
- Rebalance the architecture for future games that use more complex shaders and more memory
- Improve architectural efficiency per watt and per square millimeter
- Improve performance for DirectX 10 features such as geometry shading and stream out
- Provide significantly enhanced computation ability for high-performance CUDA™ applications and GPU physics
- Deliver improved power management capability, including a substantial reduction in idle power.

GeForce GTX 200 GPUs enable major new graphics and compute capabilities, providing the most realistic 3D graphics effects ever rendered by GPUs to date, while also providing nearly a teraflop of computational power.

Gaming Beyond: Dynamic 3D Realism

While prior-generation GPUs could deliver real-time images that appeared true-to-life in many cases, frame rates could drop to unplayable levels in complex scenes with significant animation, numerous physical effects, and multiple characters. The combination of the sheer shader processing power of GeForce GTX 200 GPUs and NVIDIA's new PhysX™ technology facilitates many new high-end graphics effects including:

- Convincing facial and character animation
- Multiple ultra-high polygon characters in complex environments
- Advanced volumetric effects (smoke, fog, mist, etc.)
- Fluid and cloth simulation
- Fully simulated physical effects such as live debris, explosions, and fires.
- Physical weather effects such as accumulating snow and water, sand storms, soaking, drying, dampening, overheating, and freezing
- Better lighting for dramatic and spectacular effect, including ambient occlusion, global illumination, soft shadows, color bleeding, indirect lighting, and accurate reflections.

Figure 1: Realistic warrior from NVIDIA "Medusa" demo

Gaming Beyond: Extreme HD

GeForce GTX 200 GPUs provide 50-100% more performance over prior-generation GPUs, permitting increased frame rates and higher visual quality settings at extreme resolutions, resulting in a truly cinematic gaming experience.

Figure 2: Far Cry 2 – Extreme HD Dynamic Beauty! (Ubisoft)

Support for the new DisplayPort interface allows resolutions beyond 2560 × 1600, and 10-bit color support permits up to a billion different colors on screen (driver, display, and application support is also required). Note that prior-generation GPUs included internal 10-bit processing, but could only output 8-bit component colors (RGB). GeForce GTX 200 GPUs permit both 10-bit internal processing and 10-bit color output.

Gaming Beyond: SLI

NVIDIA's SLI® technology is the industry's leading multi-GPU technology, giving you an easy, low-cost, high-impact performance upgrade. PC gaming simply doesn't get any faster or more realistic than running GeForce GTX 200 GPU-based boards in SLI mode on the latest nForce® motherboards.

Two flavors of SLI are supported by the initial GeForce GTX 200 GPUs:

- Standard SLI (two GPU boards), which typically boosts supported game performance by 60-90% and permits higher quality settings
- 3-way SLI, which provides even higher frame rates and permits higher quality settings for the ultimate experience in PC gaming when connected to a high-end, high-resolution monitor.

GeForce GTX 200 GPUs process and display complex DirectX 10 and OpenGL game environments with amazing graphics effects and high frame rates at extreme, high-definition resolutions.

Beyond Gaming: High-Performance Visual Computing and Professional Computation

With the power of CUDA technology and the new CUDA runtime for Windows Vista, intensive computational tasks can be offloaded from the CPU to the GPU. GeForce GTX 200 GPUs can accelerate numerous rich-media and computationally-intensive applications such as video and audio transcoding, or running distributed computing applications like Folding@home in the background while surfing the web. Examples of GPU-enabled applications include the RapidHD video transcoding application from Elemental and various video and photo editing applications.

Many engineering, scientific, medical, and financial areas demand high-performance computational horsepower for numerous applications.

Figure 3 shows the amazing speedups that can be achieved by using a GPU instead of a CPU in a number of professional visual computing applications, in addition to mainstream video transcoding. Appendix B lists references and details for these applications.

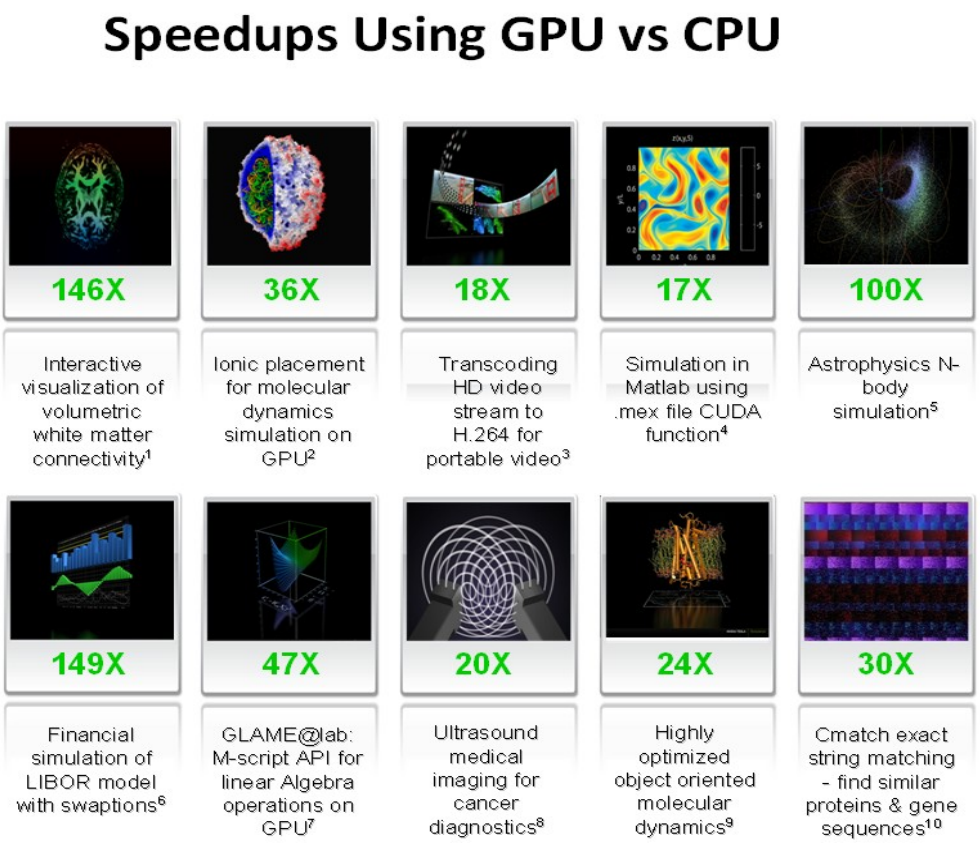

Figure 3: Significant Speedup Using GPU

With an understanding of the GeForce GTX 200 GPU design goals and key objectives, let's delve deeper into its internal architecture, looking at both the graphics and parallel processing capabilities.

GeForce GTX 200 GPU Architecture

GeForce GTX 200 GPUs are the first to implement NVIDIA's second-generation unified shader and compute architecture. The GeForce GTX 200 GPUs include significantly enhanced features and deliver, on average, 1.5× the performance of GeForce 8 or 9 Series GPUs.

Manufactured using TSMC's 65 nm fabrication process, GeForce GTX 200 GPUs include 1.4 billion transistors and are the largest, most powerful, and most complex GPU ever made. All GTX 200 GPUs are built to operate comfortably within the power and heat specifications of high-end PCs.

You may recall that the first-generation NVIDIA unified visual computing architecture in GeForce 8 and 9 Series GPUs was based on a Scalable Processor Array (SPA) framework. The second-generation architecture in GeForce GTX 200 GPUs is based on a reengineered, enhanced, and extended SPA architecture.

The SPA architecture consists of a number of TPCs, which stands for "Texture Processing Clusters" in graphics processing mode, and "Thread Processing Clusters" in parallel compute mode. Each TPC is in turn made up of a number of streaming multiprocessors (SMs), and each SM contains eight processor cores (also called streaming processors (SPs) or thread processors). Every SM also includes texture filtering processors used in graphics processing, but also useful for various filtering operations in compute mode, such as filtering images as they are zoomed in and out.

More Processor Cores

The new second-generation SPA architecture in the GeForce GTX 280 improves performance compared to the prior generation G80 and G92 designs on two levels. First, it increases the number of SMs per TPC from two to three. Second, it increases the maximum number of TPCs per chip from 8 to 10. The effect is multiplicative, resulting in 240 processor cores.

Chip	TPCs	SMs per TPC	SPs per SM	Total SPs
GeForce 8 & 9 Series	**8**	**2**	**8**	**128**
GeForce GTX 200 GPUs	**10**	**3**	**8**	**240**

Table 1: Number of GPU Processing Cores

Based on traditional processing core designs that can perform integer and floating-point math, memory operations, and logic operations, each processing core is a hardware-multithreaded processor with multiple pipeline stages that execute an instruction for each thread every clock.

Various types of threads exist, including pixel, vertex, geometry, and compute. For graphics processing, threads execute a shader program and many related threads often simultaneously execute the same shader program for greater efficiency.

All GeForce GTX 200 GPUs include a substantial portion of die area dedicated to processing, unlike CPUs where a majority of die area is dedicated to onboard cache memory. Rough estimates show 20% of the transistors of a CPU are dedicated to computation, compared to 80% of GPU transistors. GPU processing is centered on computation and throughput, where CPUs focus heavily on reducing latency and keeping their pipelines busy (high cache hit rates and efficient branch prediction).

Graphics Processing Architecture

As mentioned earlier, the GeForce GTX 200 GPUs include two different architectural personalities—graphics and computing. Figure 4 represents the GeForce 280 GTX in graphics mode. You can see the shader thread dispatch logic at the top, in addition to setup and raster units. The ten TPCs each include three SMs, and each SM has 24 processing cores for a total of 240 scalar processing cores. ROP (raster operations processors) and memory interface units are located at the bottom.

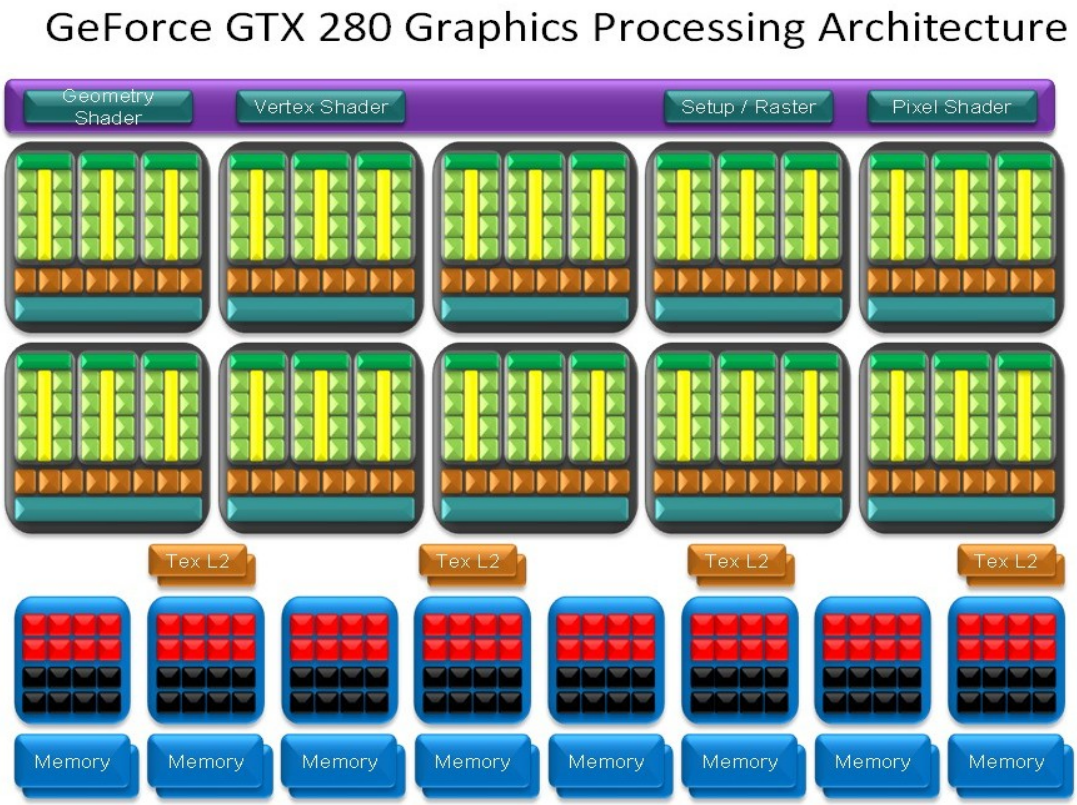

Figure 4: GeForce GTX 280 GPU Graphics Processing Architecture

Although not apparent in the above diagram, the architectural efficiency of the GeForce GTX 200 GPUs is substantially enhanced over the prior generation. We'll be discussing many areas that were improved in more detail, such as texture processing, geometry shading, dual issue, and stream out. In directed tests, GeForce GTX 200 GPUs can attain efficiencies closer to the theoretical performance limits than could prior generations.

Table 2 compares the GeForce 8800 GTX to the new GeForce GTX 280 GPU. You will notice sizable increases in a number of important measurable parameters.

Features	8800 GTX	GTX 280	% Increase
Cores	128	240	87.5 %
TEX	64t/clk	80t/clk	25 %
ROP Blend	12p/clk	32p/clk	167 %
Precision	fp32	fp64	--
GFLOPs	518	933	80 %
FB Bandwidth	86 GB	142 GB	65 %
Texture Fill	37 GT/s	48 GT/s	29.7 %
ROP Blend	7 GBL/s	19 GBL/s	171 %
PCI Express	6.4 GB	12.8 GB	100 %
Video	VP1	VP2	--

Table 2: GeForce 8800 GTX vs GeForce GTX 280

Parallel Computing Architecture

Figure 5 depicts a high-level view of the GeForce GTX 280 GPU parallel computing architecture. A hardware-based thread scheduler at the top manages scheduling threads across the TPCs. You'll also notice the compute mode includes texture caches and memory interface units. The texture caches are used to combine memory accesses for more efficient and higher bandwidth memory read/write operations. The elements indicated as "atomic" refer to the ability to perform atomic read-modify-write operations to memory. Atomic access provides granular access to memory locations and facilitates parallel reductions and parallel data structure management.

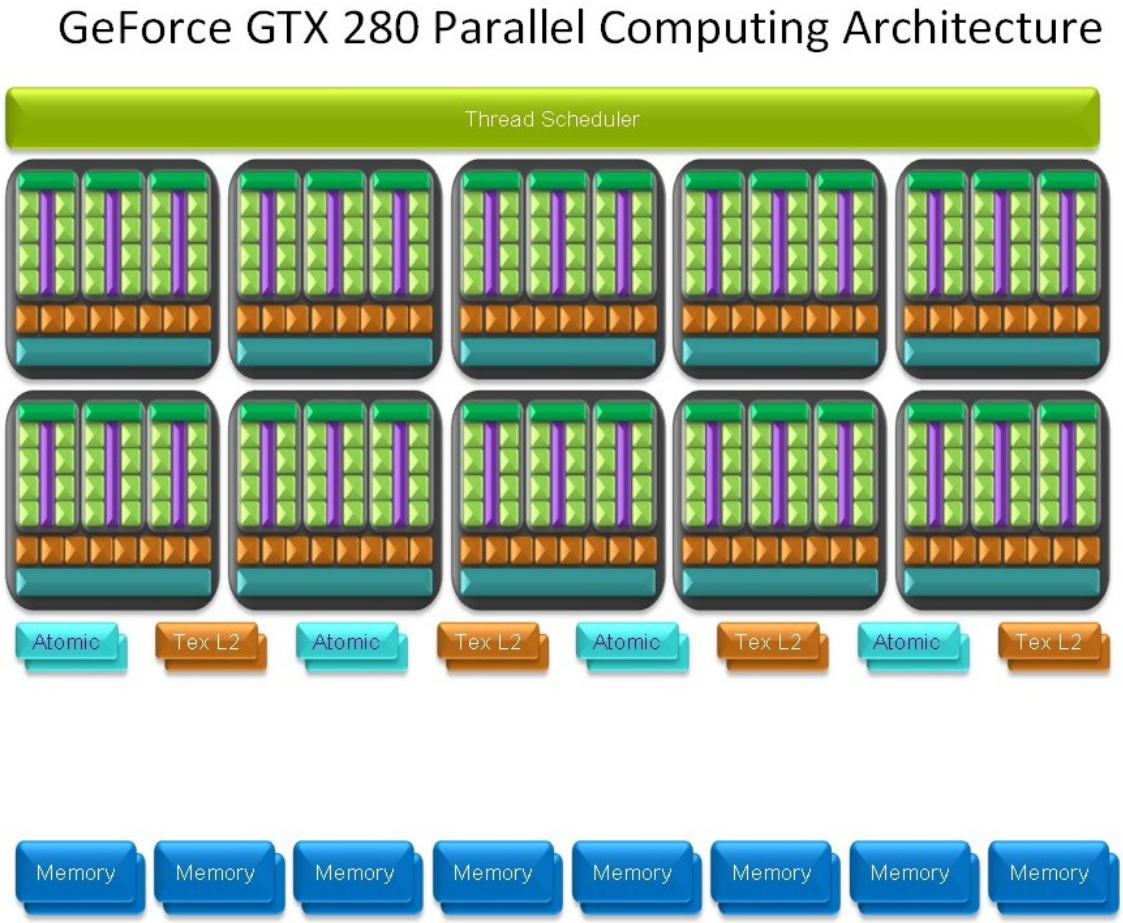

Figure 5: GeForce GTX 280 GPU Parallel Computing Architecture

A TPC in compute mode is represented in Figure 6 below. You can see local shared memory is included in each of the three SMs. Each processing core in an SM can share data with other processing cores in the SM via the shared memory, without having to read or write to or from an external memory subsystem. This contributes greatly to increased computational speed and efficiency for a variety of algorithms.

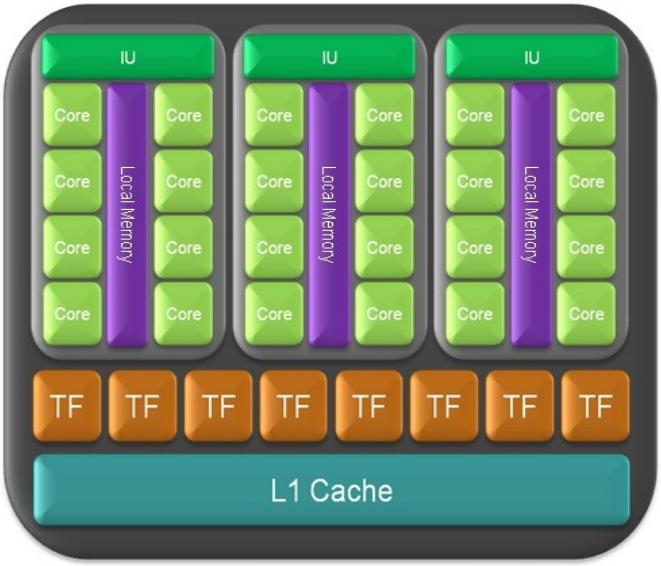

Figure 6: TPC (Thread Processing Cluster)

SIMT Architecture

NVIDIA's unified shading and compute architecture uses two different processing models. For execution across the TPCs, the architecture is MIMD (multiple instruction, multiple data). For execution across each SM, the architecture is SIMT (single instruction, multiple thread).

SIMT improves upon pure SIMD (single instruction, multiple data) designs in both performance and ease of programmability. Being scalar, SIMT has no set vector width and therefore performs at full speed irrespective of vector sizes.

In contrast, SIMD machines operate at a reduced capacity if the input is smaller than the MIMD or SIMD width. SIMT ensures the processing cores are fully utilized at all times.

From the programmer's perspective, SIMT also allows each thread to take on its own path. Since branching is handled by the hardware, there is no need to manually manage branching within the vector width.

Greater Number of Threads in Flight

GeForce GTX 200 GPUs support over thirty thousand threads in flight. Hardware thread scheduling ensures all processing cores attain nearly 100% utilization. The GPU architecture is latency-tolerant—if a particular thread is waiting for a memory access, the GPU can perform zero-cost hardware-based context switching and immediately switch to another thread to process.

The SIMT multithreaded instruction unit within an SM creates, manages, schedules, and executes threads in groups of 32 parallel threads called "warps." Up to 32 warps/SM are supported in GeForce GTX 200 GPUs, versus 24 warps/SM in GeForce 8 or 9 Series GPUs.

Chip	TPCs	SM per TPC	Threads per SM	Total Threads Per Chip
GeForce 8 & 9 Series	8	2	768	12,288
GeForce GTX 200 GPUs	10	3	1,024	30,720

Table 3: Maximum Number of Threads

Doing the math results in 32 x 32, or 1,024 maximum concurrent threads that can be managed by each SM. With 30 SMs in total, the GeForce GTX 280 supports up to 30,720 concurrent threads in hardware (versus 768 threads/SM × 2 SMs/TPC × 8 TPCs = 12,288 maximum concurrent threads in GeForce 8800 GTX).

Larger Register File

The local register file size has doubled per SM in GeForce GTX 200 GPUs compared to GeForce 8 & 9 Series GPUs. The older GPUs could run into situations with long shaders where registers would be exhausted, generating the need to swap to memory. A much larger register file permits larger and more complex shaders to be run on the GeForce GTX 200 GPUs faster and more efficiently. In terms of die size increase, the additional register file takes only a small fraction of SM die area.

Games are employing more and more complex shaders that require more register space. Figure 7 below highlights performance improvements 2× register file size in 3D Mark Vantage.

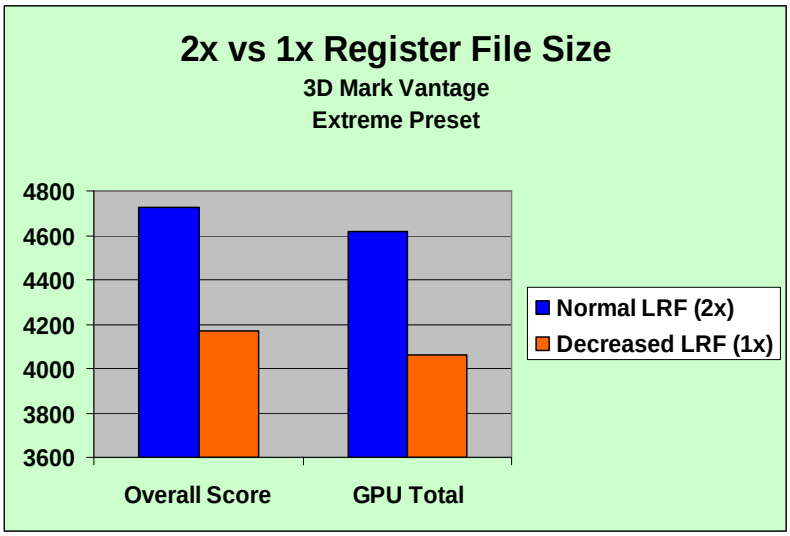

Figure 7: Local Register File 2× versus 1×

Improved Dual Issue

Special function units (SFUs) in the SMs compute transcendental math, attribute interpolation (interpreting pixel attributes from a primitive's vertex attributes), and perform floating-point MUL instructions. The individual streaming processing cores of GeForce GTX 200 GPUs can now perform near full-speed dual-issue of multiply-add operations (MADs) and MULs (3 flops/SP) by using the SP's MAD unit to perform a MUL and ADD per clock, and using the SFU to perform another MUL in the same clock. Optimized and directed tests can measure around 93-94% efficiency.

The entire GeForce GTX 200 GPU SPA delivers nearly one teraflop of peak, single-precision, IEEE 754, floating-point performance.

Double Precision Support

A very important new addition to the GeForce GTX 200 GPU architecture is double-precision, 64-bit floating point computation support. This benefits various high-end scientific, engineering, and financial computing applications or any computational task requiring very high accuracy of results. Each SM incorporates a double-precision 64-bit floating math unit, for a total of 30 double-precision 64-bit processing cores.

The double-precision unit performs a fused MAD, which is a high-precision implementation of a MAD instruction that is also fully IEEE 754R floating-point specification compliant. The overall double-precision performance of all 10 TPCs of a GeForce GTX 280 GPU is roughly equivalent to an eight-core Xeon CPU, yielding up to 78 gigaflops.

Improved Texturing Performance

The eight TPCs of the GeForce 8800 GTX allowed for 64 pixels per clock of texture filtering, 32 pixels per clock of texture addressing, 32 pixels per clock of 2× anisotropic bilinear filtering (8-bit integer), or 32-bilinear-filtered pixels per clock (8-bit integer or 16-bit floating point). Subsequent GeForce 8 and 9 Series GPUs balanced texture addressing and filtering.

- For example, the GeForce 9800 GTX can address and filter 64 pixels per clock, supporting 64-bilinear-filtered pixels per clock (8-bit integer) or 32-bilinear-filtered pixels per clock (16-bit floating point).

GeForce GTX 200 GPUs also provide balanced texture addressing and filtering and each of the 10 TPCs includes a dual-quad texture unit capable of addressing and filtering eight bilinear pixels/clock, or four 2:1 anisotropic filtered pixels/clock, or four FP16 bilinear-filtered pixels/clock. Total bilinear texture addressing and filtering capability for an entire high-end GeForce GTX 200 GPU is 80 pixels per clock.

GeForce GTX 200 GPUs employ a more efficient scheduler, allowing the chips to attain close to theoretical peak performance in texture filtering. In real world measurements, it is 22% more efficient than the GeForce 9 Series.

Chip	Theoretical Bilinear Fillrate	Measured Rate (3DMark multitex)	Measured Performance / Theoretical Performance
GeForce 9 Series	33,600	25,600	76.2%
GeForce GTX 200 GPUs	51,840	48,266	93.1%

Table 4: Theoretical vs Measured Texture Filtering Rates

Higher Shader to Texture Ratio

Because games and other visual applications are continually employing more and more complex shaders, the GeForce GTX 200 GPU design shifts the balance to a higher shader to texture ratio. By adding one more SM to each TPC, and keeping texturing hardware constant, the shader to texture ratio is increased by 50%. This shift allows the GeForce GTX 200 GPUs to perform efficiently for both today's and tomorrow's games.

ROP Improvements

The previous-generation GeForce 8 series ROP subsystem supported multisampled, supersampled, transparency adaptive, and coverage sampling antialiasing. It also supported frame buffer (FB) blending of floating-point (FP16 and FP32) render target surfaces, and either type of FP surface could be used in conjunction with multisampled antialiasing for outstanding HDR rendering quality.

The new GeForce GTX 200 GPU ROP subsystem supports all of the previous generation features, and delivers a maximum of 32 pixels per clock output, equating to 4 pixels/clock per ROP partition × 8 partitions. Up to 32 color and Z samples per clock for 8 × MSAA are supported per ROP partition. Pixels using U8 (8-bit unsigned integer) data format can be blended at twice the rate per TPC of the older-generation GPUs. Given the prior generation GPU had six ROP partitions, it could output 24 pixels/clock and blend 12 pixels/clock. In contrast the GeForce GTX 280 can output and blend 32 pixels/clock.

1 GB Framebuffer

Today's 3D games use a variety of textures to attain realism. Normal maps are used to enhance surface realism, cubemaps for reflections, and high-resolution perspective shadow maps for soft shadows. This means much more memory is needed to render a single scene than classic rendering which relied mainly on the base texture. Deferred rendering engines also make extensive use of multiple render targets, where attributes of the image are rendered off screen before the final image is composed. These techniques consume an immense amount of video memory and memory bandwidth, especially when used in conjunction with antialiasing.

The GeForce GTX 280 and GeForce GTX 260 support 1,024 MB and 896 MB of frame buffer respectively, a two-fold improvement from over prior generation GPUs. With 1 GB of frame buffer, high-resolution antialiasing performance is dramatically improved. For example, deferred rendered games like S.T.A.L.K.E.R. can now be enjoyed with antialiasing.

Geometry Shading and Stream Out

Internal output buffer structures have been significantly upsized by a factor of 6× in GeForce GTX 200 GPUs compared to the prior generation, providing much faster geometry shading and stream out performance. Figure 8 shows the latest RightMark 3D 2.0 benchmark results, including geometry shading tests. The GeForce GTX 280 GPU is significantly faster than prior generation NVIDIA GPUs and competitive products.

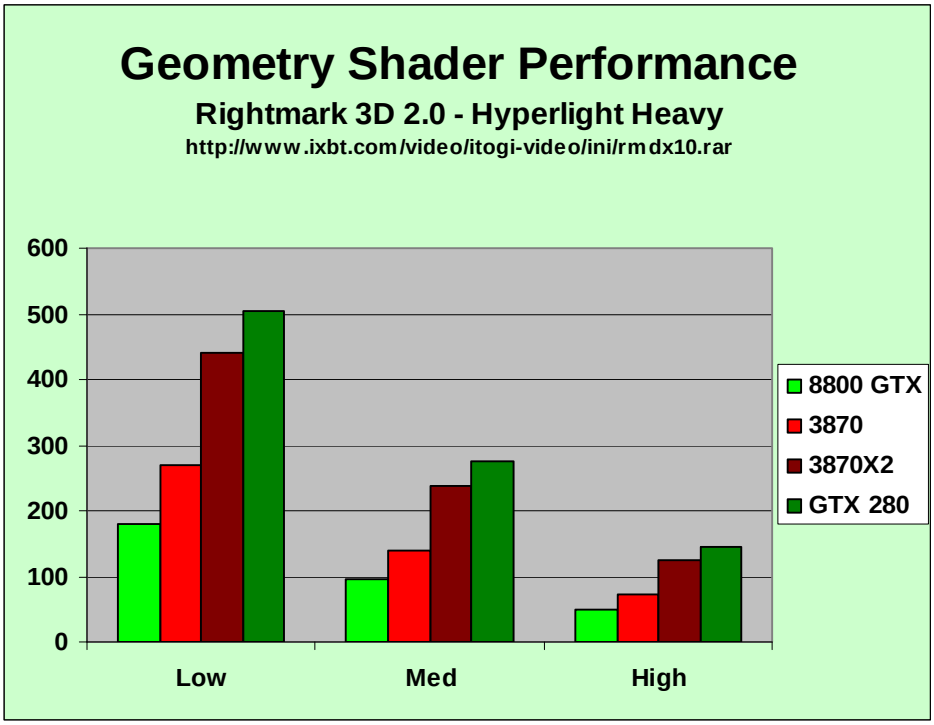

Figure 8: Geometry Shading Performance

Our own Medusa demo is highly dependent on the much faster geometry and stream out performance.

512-bit Memory Interface

Maximum memory interface width is expanded from 384 bits in previous-generation GPUs to 512 bits in GeForce GTX 200 GPUs, using eight 64-bit-wide frame buffer interface units. Memory bandwidth has been significantly increased.

In terms of rebalancing the architecture versus prior generations, the texture to frame buffer (TEX:FB) bandwidth ratio has also been modified to best support current and future workloads. NVIDIA engineers tested many applications to arrive

at the right balance of frame buffer bandwidth required to keep the texture units fully utilized and not starved for data.

General frame buffer efficiency has been improved for GeForce GTX 200 GPUs. We reworked the critical paths in the frame buffer to allow higher speed memory operation, up to 1.1 GHz GDDR3 stock speed. Memory bank access patterns and caching algorithms have also been improved. Additional compression hardware in GeForce GTX 200 GPUs effectively increase frame buffer bandwidth by permitting more data to traverse the interface per unit time, enabling better performance at higher resolutions.

Power Management Enhancements

GeForce GTX 200 GPUs include a more dynamic and flexible power management architecture than past generation NVIDIA GPUs. Four different performance / power modes are employed:

- Idle/2D power mode (approx 25 W)
- Blu-ray DVD playback mode (approx 35 W)
- Full 3D performance mode (varies—worst case TDP 236 W)
- HybridPower™ mode (effectively 0 W)

Using a HybridPower-capable nForce motherboard, such as those based on the nForce 780a chipset, a GeForce GTX 200 GPU can be fully powered off when not performing intensive graphics operations and graphics output can be handled by the motherboard GPU (mGPU).

For 3D graphics-intensive applications, the NVIDIA driver can seamlessly switch between the power modes based on utilization of the GPU. Each of the new GeForce GTX 200 GPUs integrates utilization monitors ("digital watchdogs") that constantly check the amount of traffic occurring inside of the GPU. Based on the level of utilization reported by these monitors, the GPU driver can dynamically set the appropriate performance mode (i.e., a defined clock and voltage level) that minimizes the power draw of the graphics card—all fully transparent to the end user.

The GPU also has clock-gating circuitry, which effectively "shuts down" blocks of the GPU which are not being used at a particular time (where time is measured in milliseconds), further reducing power during periods of non-peak GPU utilization.

All this enables GeForce GTX 200 graphics cards to deliver idle power that is nearly 1/10th of its maximum power (approximately 25 W on GeForce GTX 280 GPUs). This dynamic power range gives you incredible power efficiency across a full range of applications (gaming, video playback, surfing the web, etc).

Many other areas of the GeForce GTX 200 GPU pipeline have been reworked to improve performance and reduce various processing bottlenecks.

Additional Pipeline and Architecture Enhancements

Starting from the top of the GeForce GTX 200 GPUs, the front-end unit communicates with the graphics driver running on the host system to accept commands and data. The communication protocol and certain software classes have

been modified to improve efficiency of data transfer between the driver and the front end.

The memory crossbar between the data assembler and the frame buffer units has been optimized, allowing the GeForce GTX 200 GPUs to run at full speed when performing indexed primitive fetches (unlike the prior generation which suffered some contention between the front end and data assembler).

The post-transform cache size has been increased, resulting in fewer pipeline stalls and faster communication from the geometry and vertex stages to the viewport clip/cull stage. (Setup rates are similar to prior generation, supporting up to one primitive per clock).

Z-Culling performance has also been improved, especially at high resolutions. Early-Z rejection rates have been increased because the number of ZROPs was increased. The maximum ZROP cull rate is 256 samples/clock or 32 pixels/clock.

GeForce GTX 200 GPUs also include significant micro-architectural improvements in register allocation, instruction scheduling, and instruction issue. The GPUs can now feed the execution units more swiftly. These improvements are responsible for the ability to dual-issue instructions to SPs and SFUs as previously discussed. Scheduling of work between texture units and the SM controller has also been improved.

Summary

NVIDIA's second generation unified visual computing architecture as embodied in the new GeForce GTX 200 GPUs is a significant evolution over the original unified architecture of GeForce 8 and 9 series GPUs. Numerous extensions and functional enhancements to the architecture permit a performance increase averaging 1.5× the prior architecture. Improvements in sheer processing power combined with improved architectural efficiency allow amazing speedups in gaming, visual computing, and high-end computation.

Compared to earlier GPUs such as GeForce 8800 GTX, the GeForce GTX 280 provides:

- 1.88× more processing cores
- 2.5× more threads per chip
- Doubled register file size
- Double-precision floating-point support
- Much faster geometry shading
- 1 GB frame buffer with 512-bit memory interface
- More efficient instruction scheduling and instruction issue
- Higher clocked and more efficient frame buffer memory access
- Improvements in on-chip communications between various units
- Improved Z-cull and compression supporting higher performance at high resolutions, and
- 10-bit color support

These all result in enough graphics and compute power to deliver the most intensive and extreme 3D gaming experiences and teraflop performance for demanding high-end compute-intensive applications.

NVIDIA SLI technology is taken to new levels with GeForce GTX 200 GPUs and NVIDIA PhysX technology will add amazing new graphical effects to upcoming game titles. CUDA applications will benefit from additional cores, far more threads, double-precision math, and increased register file size.

Wise users purchasing new systems will conduct performance analyses to optimize their PC architecture. They will find that a lower-end CPU paired with a higher-end GPU produces more performance than the reverse and for the same price. This heterogeneous computing using the right processors for the right tasks and designing optimized PCs to take advantage of it is the wave of the future.

Appendix A: Retrospective

Over the past decade, NVIDIA's graphics processing units (GPUs) have evolved from specialized, fixed-function 3D graphics processors to highly programmable, massively multithreaded, parallel-processing architectures used for visual computing and high-performance computation.

NVIDIA GeForce GPUs enable incredibly realistic 3D gaming and outstanding high-definition video playback, while NVIDIA Quadro® GPUs provide the highest quality and fastest workstation graphics for professional design and creation. For high-performance computing tasks in various engineering, scientific, medical, and financial fields, NVIDIA's new Tesla™ GPUs and CUDA parallel programming environment enable supercomputing-level performance on the desktop, at a fraction of the cost of comparably performing CPU-based multiprocessor clusters.

The GeForce 8800 GPU was launched in November 2006. It was the world's first DirectX 10 GPU with a unified shader architecture. This was important as each of the unified shader processing cores could be dynamically allocated to vertex, pixel, and geometry workloads, making it far more efficient than prior-generation GPUs, which used a fixed number of pixel processing units and a fixed number of vertex processing units. This same unified architecture provided the framework for efficient high-end computation using NVIDIA CUDA software technology.

The GeForce 9 Series GPUs were introduced in 2007, offering a vastly improved price-performance ratio and advanced PureVideo® features. Its smaller chip allowed dual-GPU GeForce 9800 GX2 graphics boards to be built more efficiently, while offering up to twice the performance of the GeForce 8800 GTX.

As of May 2008, over 70 million NVIDIA GeForce 8 and 9 Series GPUs have shipped and each supports CUDA technology, allowing greatly accelerated performance for mainstream visual computing applications like audio and video encoding and transcoding, image processing, and photo editing. These GPUs also support the new NVIDIA PhysX technology for enabling real-time physics in games.

GPUs are the most important and most powerful processors in the new era of visual computing. High-end GeForce GTX 200 GPUs provide the best user experience when running intensive DirectX 10-based games like *Crysis* at high quality and high resolution settings. Very capable motherboard and mid-range GPUs are also needed for stutter-free, high-definition video playback on the PC while simultaneously displaying the Aero 3D user interface of Windows Vista.

Appendix B: Figure 3 References

1. "Interactive Visualization of Volumetric White Matter Connectivity in DT-MRI Using a Parallel-Hardware Hamilton-Jacobi Solver," by Won-Ki Jeong, P. Thomas Fletcher, Ran Tao, and Ross T. Whitaker

2. "GPU Acceleration of Molecular Modeling Applications."

3. Video encoding uses iTunes on the CPU and Elemental on the GPU running under Windows XP. CPUs tested were Intel Core 2 Duo 1.66 GHz and Intel Core 2 Quad Extreme 3 GHz. GPUs tested were GeForce 8800M on the Gateway P-Series FX notebook and GeForce 8800 GTS 512 MB. CPUs and GeForce 8800 GTS 512 were run on Asus P5K-V motherboard (Intel G33 based) with 2 GB DDR2 system memory. Based on an extrapolation of 1 min 50 sec 1280 × 720 high-definition movie clip.

4. http://developer.nvidia.com/object/matlab_cuda.html

5. "High performance direct gravitational N-body simulations on graphics processing units paper," communicated by E.P.J. van den Heuvel

6. "LIBOR," by Mike Giles and Su Xiaoke.

7. "FLAG@lab: An M-script API for Linear Algebra Operations on Graphics Processors."

8. http://www.techniscanmedicalsystems.com/

9. "General Purpose Molecular Dynamics Simulations Fully Implemented on Graphics Processing Units," by Joshua A. Anderson, Chris D. Lorenz, and A. Travesset

10. "Fast Exact String Matching On the GPU," presentation by Michael C. Schatz and Cole Trapnell