

Transformations

A quarterly publication from  NVIDIA.

FEATURES

Q&A with Mike Rayfield

The Visual Computing Era

3D Imaging in Stereo

Optimized PC Design

NVISION 08!



A WORD FROM MIKE

Yesterday's PC was designed as a "one-size-fits-all" system. But today's consumer is using a completely different set of applications and they are all about the visual experience. Applications like watching videos, editing photos, and playing games require optimized designs to deliver the best experience. This is visual computing and it's leading to segmentation of the PC market as system builders develop machines that outperform baseline enterprise PCs—a system for lifestyle users, another for gamers, yet another for prosumers. NVIDIA has taken a leading role in developing technology that enables visual computing applications and powers the next phase of GPU growth. In this issue we will discuss the Optimized PC design movement, our new Tegra family of computers-on-a-chip, and the growing popularity of CUDA.



Michael W. Hara
VICE PRESIDENT
INVESTOR RELATIONS AND COMMUNICATIONS

PS: We look forward to seeing you at NVISION 08, August 25-27, in San Jose, California.



with Mike Rayfield

GENERAL MANAGER, MOBILE, NVIDIA CORPORATION

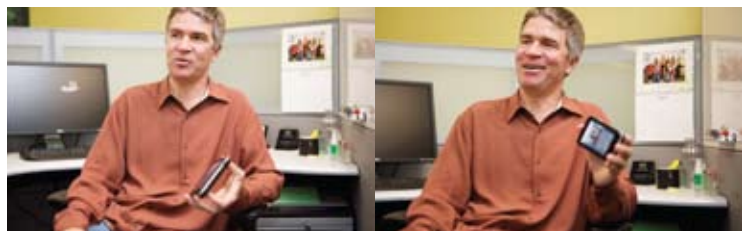
NVIDIA recently introduced Tegra, a new brand of computers-on-a-chip. What are some key points you would like us to know about Tegra?


MR: The Tegra™ products have been designed from the ground up to meet the needs of the growing universe of mobile visual computers. Smartphones, personal navigation devices, portable media players, and mobile Internet devices are all morphing into computers. They require extreme battery life and amazing multimedia performance and must fit in form factors ranging from palm size to laptops. NVIDIA's 15 years of innovation in visual computing positions us to deliver these capabilities to our customers.

Keep in mind that this is really just the beginning. Tegra will target a broad array of products, from mobile Internet devices to other delightful and engaging devices that will drive this new era of very personal computing. This is a discontinuity in the market on a scale similar to the introduction of Windows 95, which marked the start of the PC becoming a true consumer device.

How have visual mobile devices changed in the last few years?

MR: The introduction of the iPhone fundamentally changed the market in two ways. First, it raised the bar for what consumers expect from smartphones, from the quality of the audio and video to the immersive and intuitive user interface. Second, and more importantly, it marked a significant architectural change in how these devices are designed. They have evolved from being purely communications devices with limited multimedia capabilities to true multimedia computers with communications capabilities. With this shift in architecture it is easy to imagine how a phone, personal media player, personal navigation device, or mobile Internet device are all the same fundamental device with different peripheral and input/output capabilities. We've done extensive research into these and other emerging platforms and the Tegra products are tailored to deliver these subtle variations in capability.





Tegra will target a broad array of products, from mobile Internet devices to other delightful and engaging devices that will drive this new era of very personal computing.

Why did NVIDIA focus on Microsoft Windows Mobile and Windows CE?

MR: In order to build a real computer, you need a real operating system. The Microsoft operating systems deliver the necessary technical capability and enjoy widespread market adoption. We have worked with Microsoft throughout the development of Tegra and together we're ensuring that the combination of Tegra and Windows Mobile and CE will meet the challenging multimedia demands of markets around the world.

Your Tegra APX 2500 product targets a class of devices referred to as Smartphone 2.0. What does NVIDIA mean by this term?

MR: Essentially, Smartphone 2.0 is a true multimedia computer in your pocket. Of course it will still run the enterprise applications that epitomize today's smartphone, but it will be so much more. Smartphone 2.0 users will have, in the palm of their hand, an incredibly versatile device for creating and sharing content, watching and recording HD videos, listening to music, navigating, and browsing the Web.

Imagine a scenario where you are travelling to Europe for business with your APX 2500-based smartphone. You listen to music for the 12-hour flight. You check into your hotel and plug your phone into the TV via an HDMI cable and watch a two-hour HD movie. After all of this, you still have over half of your battery power available for making and receiving phone calls, browsing the Web, or sending emails. That's something that everyone can relate to and there is a clear demand in the market for technology that can deliver it.

And this isn't a scenario that is limited to the smartphone space by any means. Portable media players, mobile Internet devices, and all sorts of handheld and embedded entertainment devices are rapidly growing in capabilities beyond their core functions. Personal navigation devices, for example, now feature photo viewers, MP3 players, and Web browsers on top of their core navigation capabilities.

The NVIDIA Tegra products deliver the technology needed to enable high-performance and engaging applications that merely sip at the battery without sacrificing quality. It is this kind of usability that truly embodies the next generation of personal computers and NVIDIA, with our proven expertise in visual computing, is uniquely positioned to deliver it. ☺





The Visual Computing Era

Spotlight on CUDA and Scalable Parallel Programming

BY JOHN NICKOLLS, IAN BUCK, KEVIN SKADRON (ON SABBATICAL FROM UNIV. OF VIRGINIA) AND MICHAEL GARLAND, NVIDIA.
ADAPTED FROM AN ARTICLE THAT APPEARED IN ACM QUEUE MAGAZINE, APRIL 2008

The advent of multi-core CPUs and many-core GPUs means that mainstream processor chips are now parallel systems. Furthermore, their parallelism continues to scale with Moore's Law. The challenge is to develop mainstream application software that transparently scales its parallelism to leverage the increasing number of processor cores, much as 3D graphics applications transparently scale their parallelism to many-core GPUs with widely varying numbers of cores.

According to conventional wisdom, parallel programming is difficult. However, early experience with the CUDA™ scalable parallel programming model and C language shows that many sophisticated programs can be readily expressed with a few easily understood abstractions. Since NVIDIA released CUDA technology in 2007, developers have rapidly developed scalable parallel programs for a wide range of applications, including financial modeling, computational chemistry, and oil & gas exploration. These applications scale transparently to hundreds of processor cores and thousands of concurrent threads.

NVIDIA GPUs with the new Tesla™ unified graphics and computing architecture run CUDA C programs and are widely available in laptops, PCs, workstations, and servers. The CUDA model is also applicable to other shared-memory parallel processing architectures, including multi-core CPUs.

Unified Graphics and Computing GPUs

Driven by the insatiable market demand for real-time, high-definition 3D graphics, the programmable GPU has evolved into a highly-parallel, multithreaded, many-core processor. It is designed to efficiently support the graphics shader programming model, in which a program for one thread draws one vertex or shades one pixel fragment. The GPU excels at fine-grained, data-parallel workloads consisting of thousands of independent threads executing vertex, geometry, and pixel shader program threads concurrently.

The tremendous raw performance of modern GPUs has led researchers to explore mapping more general non-graphics computations onto the GPU processors. Such GPGPU—or General-Purpose Computation on GPUs—systems have produced some impressive results, but the limitations and difficulties of doing this via graphics APIs are legend. This broad-based desire to use the GPU as a more general parallel computing device motivated NVIDIA to develop a new unified graphics and computing GPU architecture and the CUDA programming model.



GPU Computing Architecture

Introduced by NVIDIA in November 2006, the Tesla unified graphics and computing architecture significantly extends the GPU beyond graphics—its massively multithreaded processor array becomes a highly-efficient unified platform for both graphics and general-purpose parallel-computing applications. By scaling the number of processors and memory partitions, the Tesla architecture spans a wide market range from the high-performance enthusiast GeForce® GPUs and professional Quadro® and Tesla computing products to a variety of inexpensive, mainstream GeForce GPUs. Its computing features enable straightforward programming of the GPU processor cores in C with CUDA. Wide availability in laptops, desktops, workstations, and servers, coupled with C programmability and CUDA software, make the Tesla architecture the first ubiquitous supercomputing platform.

CUDA Paradigm

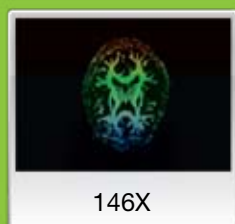
CUDA provides three key abstractions—a hierarchy of thread groups, shared memories, and barrier synchronization—that provide a clear parallel structure to conventional C code for one thread of the hierarchy. Multiple levels of threads provide fine-grained parallelism and coarse-grained parallelism. The abstractions guide the programmer to partition the problem into coarse sub-problems that can be solved independently in parallel, and then into finer pieces that can be solved in parallel. The programming model scales transparently to large numbers of threads and processor cores: a compiled CUDA program executes on any number of processors, and only the run time system needs to know the physical processor count.

CUDA Application Experience

CUDA minimally extends the C and C++ programming languages. The programmer writes a serial program that calls parallel kernels, and writes serial C code for the kernels, which may be simple functions or full programs. A kernel executes in parallel across a set of parallel threads. Programmers who are comfortable developing in C can begin writing CUDA programs in a very short amount of time.

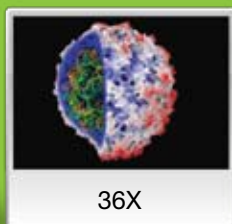
In the relatively short period since its introduction, a number of real-world parallel application codes have been developed in CUDA. These include financial modeling, FHD spiral MRI reconstruction, molecular dynamics, and n-body astrophysics simulation.

Continued on following page>>



146X

Interactive visualization of volumetric white matter connectivity¹



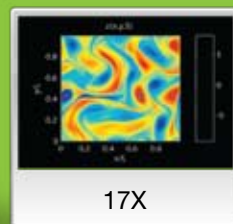
36X

Ionic placement for molecular dynamics simulation on GPU²



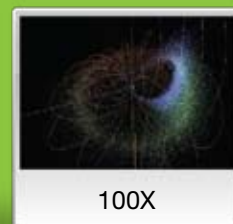
19X

Transcoding HD video stream to H264 for portable video³



17X

Simulation in Matlab using .mex file CUDA function⁴



100X

Astrophysics N-body simulation⁵

Examples of speedup results using CUDA compared to previous approaches.

Running on Tesla-architecture GPUs, these applications were able to achieve substantial speed-ups over alternative implementations running on serial CPUs: the financial modeling was 149 times faster, the MRI reconstruction was 263 times faster, the molecular dynamics code was 10–100 times faster, and the n-body simulation was 50–250 times faster. The large speedups are due to the highly parallel nature of the Tesla architecture and its high memory bandwidth.

Conclusion

CUDA is a model for parallel programming that provides a few easily understood abstractions that allow the programmer to focus on algorithmic efficiency and develop scalable parallel applications. In fact, CUDA is an excellent programming environment for teaching parallel programming. The University of Virginia used it as a short, three-week module in an undergraduate computer architecture course, and students were able to write a correct k-means clustering program after just three lectures.

The University of Illinois has successfully taught a semester-long parallel programming course using CUDA to a mix of Computer Science and non-Computer Science students, with students obtaining impressive speedups on a wide variety of real applications, including the previously-mentioned MRI reconstruction example.

The programming paradigm provided by CUDA has allowed developers to harness the power of these scalable parallel processors with relative ease, enabling them to achieve speedups of 100 times or more on a variety of sophisticated applications. The CUDA abstractions, however, are general and also provide an excellent programming environment for multi-core CPU chips. A prototype source-to-source translation framework developed at the University of Illinois compiles CUDA programs for multi-core CPUs by mapping a parallel thread block to loops within a single physical thread. CUDA kernels compiled in this way exhibit excellent performance and scalability.

Although CUDA was released not that long ago, it is already the target of massive development activity—there are tens of thousands of CUDA developers. The combination of massive speedups, an intuitive programming environment, and affordable, ubiquitous hardware are unique in today's market, and represent the realization of decades of hopes in parallel computing. In short, we believe CUDA represents a democratization of parallel programming.

Links to the latest version of the CUDA development tools, documentation, code samples, and user discussion forums can be found at: www.nvidia.com/CUDA.



149X

Financial simulation of LIBOR model with swaptions⁶



47X

GLAME@lab: An M-script API for linear algebra operations on GPU⁷



20X

Ultrasound medical imaging for cancer diagnostics⁸



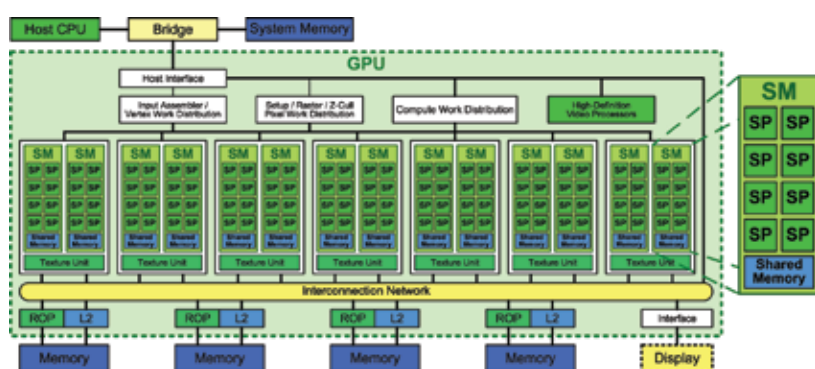
24X

Highly optimized object oriented molecular dynamics⁹



30X

Cmatch exact string matching to find similar proteins and gene sequences¹⁰



NVIDIA Tesla Architecture GPU with 112 Streaming Processor (SP) Cores, arranged as 14 Streaming Multiprocessors (SMs).

HEADLINE IMAGERY SOURCES:

- 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, IEEE Conference on Visualization, Oct. 2007. <http://www.cs.utah.edu/%7Ewkjeong/publication/vis07.pdf>.
- 2 "Accelerating molecular modeling applications with graphics processors", by John E. Stone, James C. Phillips, Peter L. Freddolino, David J. Hardy, Leonardo G. Trabuco, Klaus Schulten. Journal of Computational Chemistry, vol. 28, no. 16, 2007, pp 2618–2640, <http://dx.doi.org/10.1002/jcc.20829>.
- 3 Video encoding test uses iTunes on CPU and Elemental RapiHD on GPU running under Windows XP on Intel Core 2 Duo 1.66GHz and Intel Core 2 Quad Extreme 3GHz. GPUs were GeForce 8800M on Gateway P-Series FX notebook, and GeForce 8800 GTS 512MB on Asus P5K-V motherboard (Intel G33-based) with 2GB DDR2 system memory. Extrapolated from 1 min 50 sec 1280x720 HD movie clip. <http://www.elementaltechnologies.com/>
- 4 "MATLAB plug-in for CUDA". http://developer.nvidia.com/object/matlab_cuda.html.
- 5 "High-Performance Direct Gravitational N-body Simulations on Graphics Processing Units—II: An Implementation in CUDA", Robert G. Belleman, Jeroen Bedorf, Simon Portegies Zwart, New Astronomy, July 2007, arXiv:0707.0438v2 <http://arxiv.org/abs/0707.0438v2>
- 6 LIBOR Monte Carlo code and "Notes on using the NVIDIA 8800 GTX graphics card", by Mike Giles and Su Xiaoke. <http://www2.maths.ox.ac.uk/~gilesm/hpc/NVIDIA/libor/report.pdf>
- 7 "GLAME@lab: An M-script API for Linear Algebra Operations on Graphics Processors", Sergio Barrachina, Maribel Castillo, Francisco D. Igual, Rafael Mayo, Enrique S. Quintana-Ort, Feb. 2008, <http://www3.uji.es/~figual/files/Papers/para08.pdf>.
- 8 See <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, Journal of Computational Physics, vol 227, no. 10, May 2008, DOI: 10.1016/j.jcp.2008.01.047, <http://www.external.ameslab.gov/hoomd/downloads/GPUMD.pdf>
- 10 "Fast Exact String Matching on the GPU", Michael C. Schatz and Cole Trapnell, May 2008, <http://www.cbcb.umd.edu/software/cmatch/Cmatch.pdf>

Stanford University Launches New Pervasive Parallelism Lab

Last month, NVIDIA announced that it is a founding member of Stanford University's new Pervasive Parallelism Lab (PPL). The PPL will develop new techniques, tools, and training materials to allow software engineers to harness the parallelism of the multiple processors that are already available in virtually every new computer. NVIDIA's investment complements the company's ongoing strategy to solve some of the world's most computationally-intensive problems with GPUs and world-class tools and software.

Until recently, computer installations delivering massive parallelism could be deployed only in large-scale computer centers with hundreds to thousands of separate computer systems. With the recent introduction of many-core processors such as the GPU and the multi-core CPU, most new computer systems come equipped with multiple processors that require new software techniques to exploit parallelism. Without new software techniques, computer scientists are concerned that rapid increases in the speed of computing could stall.

From fundamental hardware to new user-friendly programming languages that will allow developers to exploit parallelism automatically, the PPL will allow programmers to implement their algorithms in accessible, "domain-specific" languages while at deeper, more fundamental levels of software, the system would do all the work for them in optimizing the code for parallel processing.

NVIDIA is joined by AMD, HP, IBM, Intel, and Sun in this venture.

3D Imaging in Stereo

An intriguing area of imaging technology is the field of research known as “naked-eye stereoscopy,” which displays 3D stereoscopic images without the need for special eyeglasses. This technology not only has applications in entertainment, but is being studied as a practical technology for a variety of professional applications.

One especially promising application is in medical imaging, where NVIDIA's CUDA parallel computing platform is being studied by Professor Takeyoshi Dohi and his colleagues in the Department of Mechano-Informatics at the University of Tokyo's Graduate School of Information Science and Technology.

Naked-eye stereoscopy can be implemented in a variety of ways; the one being studied by Associate Professor Hongen Liao and graduate student Nicholas Herlambang in Professor Dohi's research group is called Integral Videography (IV). This method uses a special display comprised of a micro-lens array with convex lenses on a matrix which is bonded to a liquid crystal panel. Directly beneath each micro-lens, there are some 100 liquid-crystal elements. The convex lens projects the light from each element in various directions. The object to be represented in 3D space is illuminated by light rays from several directions, forming a stereoscopic image which, to the user, seems to be floating in the air.

Because this method projects a 3D image into space, it has advantages over the traditional stereoscopic method, where different images are displayed for the viewer's left and right eyes. Using IV, the 3D image can be observed from a wide area in front of the display by several viewers at once, without using special eyeglasses or viewpoint tracking.

Since 2000, the university's research group has been developing a system where in vivo cross-sections obtained in real time by CT or MRI scans are treated as volume textures, which can not only be reconstituted as 3D images through volume rendering, but also displayed as stereoscopic video for use in an IV system.

This system could revolutionize real-time, stereoscopic, in-vivo imaging. However, the amount of computation is huge; the volume rendering alone creates a high processing load, then further processing is required for the stereoscopic imaging. For each video frame, a vast number of angles must be displayed at the same time. Multiply this by the number of frames in the video and a staggering amount of computation must be done with high precision in a short time.

In research in 2001, real-time volume rendering and stereoscopic reconstitution for images of 512 x 512 resolution on a Pentium III 800 MHz PC took over ten seconds to generate a single frame. To speed up processing, the group tried using the 60 CPUs of an UltraSPARC III 900 MHz machine, the latest high-performance computer available at the time. But the best result that could be obtained was five frames

About the ATRE Lab



The Advanced Therapeutic and Rehabilitation Engineering (ATRE) Laboratory is a multi-disciplinary research center working on computer-aided surgery and rehabilitation engineering. Its focus is on the development of robotic devices and computer programs to assist physicians with minimally invasive surgery, as well robotically-enhanced assisting devices for the elderly and physically challenged. The laboratory is part of the Department of Mechano-Informatics in the Graduate School of Information Science and Technology at the University of Tokyo.

LEFT TO RIGHT Professor Takeyoshi Dohi, Associate Professor Hongen Liao, Graduate Student Nicholas Herlambang

per second. This was simply not fast enough to be practical.

Both the volume rendering and subsequent conversion to IV format require data-parallel vector calculation. For this, the optimal computing paradigm is the GPU. Accordingly, Liao and Herlambang started to research GPU implementation using CUDA, a general-purpose, C-language GPU development environment from NVIDIA.

First, the researchers developed a prototype system based on the NVIDIA architecture. When data sets from the 2001 study were run on the GPU using CUDA, performance improved to 13-14 frames per second. As the UltraSPARC system had cost tens of millions of yen, the researchers were amazed that a GPU, costing a hundred times less, delivered nearly three times the performance. Moreover, according to the group, NVIDIA's GPU was at least 70 times faster than the latest generation of multi-core CPUs. In addition, tests showed that the GPU's high performance was even more conspicuous for larger volume texture sizes.

Currently, the group is working with the Tesla D870-based deskside supercomputer and is optimizing

the current IV system for Tesla using CUDA. This is expected to boost performance even further.

"We have evaluated various development environments for parallel computing," says Herlambang, the researcher responsible for the CUDA implementation, "but we chose CUDA because it lets us do development in the C language syntax we are used to. Also, we will be able to take advantage of speed increases in future generations of GPUs without modifying the system we have developed. If an environment that makes it easy to debug large CUDA programs becomes available, CUDA will become an even more powerful development environment for parallel computing and we expect it will find more applications in medical image processing as well."

When images from CT and MRI are viewed stereoscopically in real time, physicians can check the state of diseased tissues and make diagnoses without biopsies and surgery. Moreover, several physicians can view the images at the same time and consult with one another. And it may eventually make it possible for several physicians to perform arthroscopic surgery and other minimally invasive surgical techniques together, with

each surgeon able to visualize the operation in real time.

It is difficult to bring a huge parallel computer array into clinical settings, but the powerful computing capabilities of GPUs make it possible to provide compact, parallel computing modules. @

Kaori Nakamura and David Wilton of NVIDIA contributed to this article.

IV system using CUDA.




Optimized PC Design

The User Experience is Paramount

Consumers are demanding PCs that deliver richer graphics and beautiful visuals. To respond, PC manufacturers all over the world are no longer designing PCs based simply on the “speeds and feeds” of the CPU, but with the total user experience in mind.

A PC should be designed and optimized for how the user intends to use it. The more visual the experience, the more important the investment in the GPU becomes. And for large segments of the marketplace like workstations, gaming PCs, video/photo editing PCs, media centers, and lifestyle PCs, the visual experience is front-and-center and cannot be compromised. The GPU is no longer a luxury or merely a “nice-to-have.”

All over the world, we see the movement toward usage-optimized PC design taking hold and increasing in momentum. We call this the “Optimized PC” design movement. This trend, combined with growing numbers of visually-rich applications and receptive consumers with high expectations, is driving consumption of GPUs. 

An Optimized PC Improves the Visual Computing Experience



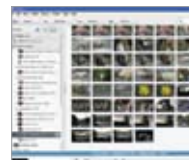
Games

World in Conflict™, ©2008
Massive Entertainment AB.



Video

Golden Gate Bridge
PureVideo® HD technology



Photos

PicLens ©2008
Cooliris, Inc.



Operating Systems

Microsoft® Windows Vista™
©2008 Microsoft

nVISION 08
THE WORLD OF VISUAL COMPUTING

NVISION 08!

The First-Ever Visual Computing Mega-Event

PLEASE JOIN US AT NVISION 08, AUGUST 25-27, 2008 IN SAN JOSE, CALIFORNIA

Featuring industry luminaries, in-depth technical sessions, cutting-edge technology previews, and live musical and visual entertainment, NVISION 08 will focus on the intersection of technology, science, and art that defines the world of visual computing.

The event will be held August 25-27, 2008 in San Jose, California, promising to turn Silicon Valley into a mecca of visual computing this summer. Conference registration and sponsorship information is available at www.nvision2008.com.

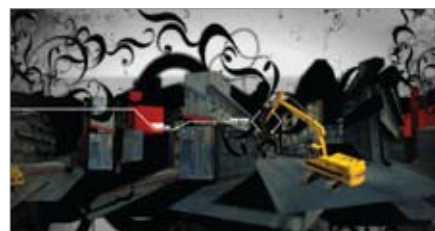
For more information, contact Calisa Cole at ccole@nvidia.com.

“NVISION 08 will be the defining event for visual computing. It brings the visual computing ecosystem under one roof to explore, share, and discover.”

-Jen-Hsun Huang, CEO, NVIDIA



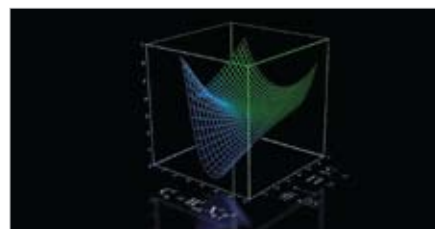
Professional Visualization



Demoscene Community



Digital Art



Computing

Transformations is a quarterly publication
of NVIDIA's investor relations and communications group.

Please send feedback to nvtransformations@nvidia.com.

NVIDIA Corporation | 2701 San Tomas Expressway | Santa Clara, CA 95050 | www.nvidia.com

Copyright © 2008 NVIDIA Corporation. All rights reserved. NVIDIA, the NVIDIA logo, GeForce, Quadro, Tesla, NVIDIA Tegra, CUDA, PureVideo and NVISION are trademarks and/or registered trademarks of NVIDIA Corporation in the U.S. and other countries. Other company and product names may be trademarks of the respective companies with which they are associated.

