issue 03 // fall-winter 07

Transformations

A quarterly publication from NVIDIA.

FEATURES

A WORD FROM MIKE

NVIDIA will celebrate its 15th year anniversary in February. As I reflect on the accomplishments and breakthroughs that we've made during this time, a characteristic of NVIDIA that stands out is our ability to innovate at an absolutely relentless pace, quarter after quarter, year after year. This passion for pushing the boundaries of technology and imagination drives all of our businesses—from consumer products to professional solutions—and positions us firmly as a leader in the visual computing revolution. Developers, scientists, engineers, and researchers around the world are finding new and exciting ways to leverage the GPU across a wide range of areas, from film making to medical diagnosis to oil & gas exploration. From our vantage point, we see consumers with an insatiable appetite for as much processing power as we can offer, making NVIDIA a true "friend of Moore's Law." As I contemplate the potential of the GPU to transform industries and solve the world's most difficult problems, I am excited to think about what the next 15 years will hold not only for NVIDIA but for our entire worldwide ecosystem of partners, developers, and consumers.

Michael W. Hara
VICE PRESIDENT
INVESTOR RELATIONS AND COMMUNICATIONS

Q&A

with Marv Burkett

CFO, NVIDIA CORPORATION

GPUs are one of the few areas that can efficiently use the doubling of transistors.

Describe some of the changes the semiconductor industry has undergone over the past decade.

MB: I see two significant changes in the industry that are evident today. The first is the entrance of private equity into the semiconductor industry. The acquisition of On Semiconductor, Freescale, and Phillips Semiconductor, all by private equity groups, signals a change in the perception of the business. Private equity has been around for a long time, but until recently they have been unwilling to invest in the semiconductor business. The primary reason for that is that in the past, semiconductor businesses were consumers of cash, not cash generators. If private equity uses significant debt to finance their purchases, then they are unwilling to invest in cash consumers because they need to be able to service the debt. That has changed in the last ten years. Now most semiconductor companies are generating cash because they are outsourcing the fab and/or sharing R&D costs. This attracts private equity and impacts the valuations of semiconductor companies.

The second change is the disappearance of the IDMs (Integrated Device Manufacturers). In the 70's and 80's most semiconductor companies were IDMs, meaning they had their own fabs. In the 80's we saw the start of the fabless business model and the emergence of pure foundries. The prohibitive cost of fabs, coupled with the skyrocketing cost of process development, has forced many IDMs to abandon the fab strategy. The announcement by TI that they were no longer going to develop process technology was a sea change in the industry. In the U.S. that leaves only Intel and IBM as leading-edge process developers and will eventually leave only Intel as an IDM. The emergence of companies like TSMC with significant leading edge fab capacity has not only allowed this change, but forced it, with very cost effective capacity. There is no longer an advantage to owning your fab.

Talk about the relationship between NVIDIA and the investment community.

MB: I believe investors have a much better understanding about NVIDIA than they did a few years ago. Previously there was the perception that graphics suppliers traded position every cycle. This wasn't true, but it was a common perception with investors. Now they understand that with few exceptions, NVIDIA has been and continues to be the technology leader.

The other perception that has changed is the inherent gross margins in our business. For a while, when we were improving gross margins, there was the perception that soon gross margins would fall and return to the old levels. It is only with our continued progress and an understanding of inherent causes of the change, that they have changed their minds. Now they believe we have fundamentally changed the business model for revenue per wafer, which changes their view of our inherent gross margins. Investors are interested in earnings growth. What excites them about NVIDIA is that we have the potential for both a.) revenue growth, which can lead to earnings growth, and b.) expansion of gross margins, which will also lead to earnings growth.

Does NVIDIA view Moore's Law as a friend or foe? Please explain.

MB: Moore's Law has lived longer than most expected. It's over 40 years old and is still alive and well. At this point, each doubling of the # of possible transistors creates significant challenges and opportunities. We are at the point where each successive generation of technology adds not thousands or even millions of possible transistors, but now a new generation could add a billion transistors or more. The industries or companies that can use a billion more transistors are dwindling. Memories can always use more density, so Moore's Law would be their friend. CPUs may be able to use the additional transistors, but can they do it efficiently? That is, is a dual core processor twice as powerful as a single core? Is a quad core four times as powerful and useful as a single core? Probably not. Therefore I think it can be argued that Moore's Law is not the friend of CPU companies. GPUs are one of the few areas that can efficiently use the doubling of transistors. This means that with a doubling of transistors, GPU designers can double the performance. How long this will go on, we don't know, but for the foreseeable future, GPUs have ways of doubling the performance with successive generations. So certainly Moore's Law is NVIDIA's friend.

TRANSFORMATIONS // NVIDIA CORPORATION // WWW.NVIDIA.COM

The Future of 3D Graphics

Spotlight on GPU Computing and Ray Tracing

THE FIRST IN A SERIES OF ARTICLES ABOUT 3D GRAPHICS BY NVIDIA'S CHIEF SCIENTIST, DAVID KIRK

GPUs have evolved far beyond simply implementing a fixed function graphics pipeline to becoming flexible, programmable, massively parallel computers.

In recent years we've seen tremendous interest in utilizing the immense parallel processing power of GPUs for uses beyond classic 3D graphics processing. GPUs have evolved far beyond simply implementing a fixed function graphics pipeline to becoming flexible, programmable, massively parallel computers. Similarly, 3D graphics has evolved to encompass many forms of visual computing applications. GPUs are now considered "computational graphics" engines, as many of the fixed function parts of the graphics pipeline have become programmable. And we have only seen the beginning of this transformation.

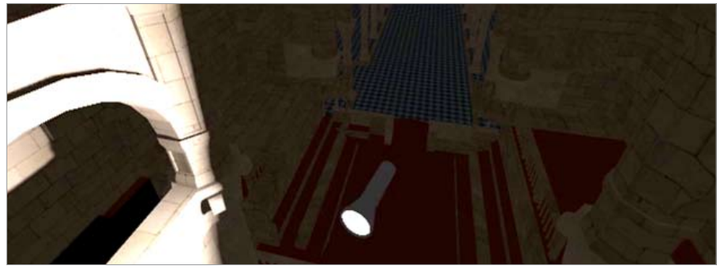


FIGURE 1

GPUs Lead Evolution to Many-Core Processing

The programmable and flexible modern GPU is one of the most powerful computing devices on the planet. Since the year 2000, the individual processing cores within GPUs have processed data using IEEE floating-point precision, just like standard CPUs (aka "real computers"). The raw floating-point processing power of a modern GPU is much larger and growing faster than even the latest multi-core CPU. This feature has attracted a lot of attention in the computing community. In fact, an entirely new field of effort has been spawned called GPGPU, or General Purpose Processing on GPUs, reflecting the desire to use the power of the GPU for broader applications than just graphics. More recently, this broader effort, which we call GPU Computing, has been made easier by the introduction of NVIDIA's CUDA (Compute Unified Device Architecture) programming environment. CUDA allows GPUs to be programmed using the C language for non-graphics applications.

All processors evolve and change with time, not just GPUs. We are seeing this with the difficult transition that CPUs are making from single-core to multi-core. Due to problems with heat dissipation and power consumption, it is no longer possible to create faster CPUs that simply run

FIGURE 2

at ever higher clock rates. However, it's quite easy to add multiple CPU cores to a single chip—but that's where the simplicity ends. It is difficult for programmers to grasp how to program multi-core CPUs effectively. Also, for the first time in several decades, programmers can no longer simply wait 18-24 months for their single-threaded programs to double in speed as processor clock speeds increase. An industry wide effort to "refactor" algorithms to run on multi-core CPUs is taking place while, at the same time, the emergence of GPU Computing gives programmers a new powerful tool. Today, few application programs benefit from multi-core CPUs. In contrast, the GPU programming model aided by graphics APIs and the C programming language is straightforward and easy to use. Although applying GPUs to a variety of parallel computing tasks is a natural evolution, trying to process demonstrably parallel graphics workloads with multi-core CPUs is inherently challenging—because simply grouping together many CPUs will not produce an integrated parallel processor. A GPU consists of many parallel processor cores integrated to work together from the ground up.

Graphics processors have arguably been multi-core processors for almost ten years. In 1998, NVIDIA's TNT product was built with two pixel pipelines and two texture mapping units. We never looked back—the current GeForce 8800 chip has 128 processor cores. And, not only does it have 128 processor cores, but each core can run many threads, or program copies, at a time. The GeForce 8800 processes over 12,000 threads at once—each thread processing pixels, vertices, or triangles! Imagine achieving that kind of parallelism and throughput with dual-or quad-core CPUs. It's just not possible. But, that's not all. In addition to the 12,000+ pixel or vertex threads, there are many thousands of other concurrent operations being processed by the GPU. Texture map calculations, rasterization, Z-buffer hidden-surface-removal, color blending for transparency, and anti-aliasing (edge smoothing) are all happening simultaneously. Without the special-purpose hardware included in every GPU to perform these operations, it would require hundreds if not thousands of CPU cores to match the performance of a single GPU.

Continued on following page>>

TRANSFORMATIONS // NVIDIA CORPORATION // WWW.NVIDIA.COM

We believe the most valuable architectures are those that extend rather than disrupt the installed base, such as x86, Windows, HTML, and TCP/IP. Preserving the industry's investments is important not only for maintaining productivity of current applications and developers but also for encouraging investment in future applications.

Is Ray Tracing Ready for Mainstream Use?

Some recent blogs and press reports have commented that the future of 3D graphics will be based on the feature known as "ray tracing," and therefore the performance of rasterization is unimportant. While we are enormous fans of interactive ray tracing (IRT), it still requires a massive amount of processing power. IRT, along with many other ideas we are pursuing, will provide at least another decade of innovation opportunity for GPU designers. The sustainable innovation opportunity will keep the GPU industry vibrant. But rather than a "start from scratch" architectural approach, we believe we need to preserve the massive investments of all the industries that are deeply invested in OpenGL and DirectX. We believe the most valuable architectures are those that extend rather than disrupt the installed base, such as x86, Windows, HTML, and TCP/IP. Preserving these investments is important not only for maintaining the productivity of current applications and developers, but also for encouraging investment in future applications.

Furthermore, ray tracing is not a panacea or really a goal in itself, but rather—potentially—a way to make better pictures more easily. Although some people may say that ray tracing is more accurate or "the right way," both ray tracing and rasterization are approximations of the physical phenomena of light reflection from surfaces. Neither is inherently better or worse—just different. Three possible reasons for adopting ray tracing are ease of programming, faster visual effect, and the possibility of better visual effects. Let's talk about these reasons separately.

Ease of programming is important for 3D graphics as well as for many other applications. Ray tracing is believed to be "easier" for programmers than rasterization because ray tracing can do everything in a single unified approach. Although it is true that rays can be traced for every possible visual effect, this is not necessarily the best and fastest approach.

Ray tracing will never be as fast as hardware rasterization for the purpose of visibility (i.e., which objects the eye sees directly). Simple visibility is not enough, however. We also would like anti-aliasing (i.e., the smoothing of edges). Ray tracing can accomplish this effect by simply tracing more rays, although this is more expensive and slower than allowing the GPU hardware to perform this task through rasterization and dedicated anti-aliasing hardware.

One visual effect that is difficult to do well with rasterization is shadows. It is complicated to render sharp-edged shadows without having jagged edges, and there are no really robust approaches for making soft-edge shadows or the corresponding effects of multiple inter-reflections of light. These are visual effects for which ray tracing is indeed a more general solution. Also, as light reflects from each surface to shine on other surfaces, every single object in a scene is both a light source and an occluder (an object that blocks light). In order to make the "perfect" picture, you would need to trace a ray from every point on every object in every direction. A lot of rays would be required to simulate all of that light bouncing around. Although this is conceptually simple, it is more work than is practical with modern CPUs.

Parallel GPU hardware can trace rays as well. Making a picture that is noticeably better than what rasterization can do today is a difficult but worthy goal. That is one of the most exciting parts of the field of computer graphics—we're never done. There is always something more to accomplish. The picture of the cathedral in this article is shown in two versions. *Figure 1* could be rendered either with ray tracing or rasterization. There is no special effort involved in making shadows or correct global illumination. *Figure 2* is rendered using global illumination. This is the effect of light inter-reflecting between surfaces. The back of the cathedral—behind the light—is lit by the reflection of the light off of the surfaces in front of the light. By the way, both of these images are rendered using a GPU. The bottom line is that a GPU can now make any picture that a CPU can make, rasterized or ray traced. The combination of special purpose GPU hardware and APIs (DirectX and OpenGL) with computational graphics is powerful.

The GPU Can Do It All

The old debate of ray tracing vs. rasterization has been going on as long as there has been ray tracing and rasterization. The debate has now morphed into ray tracing vs. GPUs. But, comparing an approach (ray tracing) with a device (GPU) is a funny way to express the comparison. That's like asking which is better, fuel or automobiles? Most likely, the answer to both questions is both. Not only do GPUs perform rasterization efficiently using the conventional API-based programmable graphics pipeline, but GPU computing has the promise of performing other rendering approaches as well. It is likely that game developers, film studios, animators, and artists will prefer to take advantage of all of the benefits of ray tracing and rasterization, as well as a variety of other techniques, all at the same time. Why choose when you can have it all? Interestingly, one of the best GPGPU applications may be…3D graphics!

Nick Stam and Kevin Krewell of NVIDIA contributed to this article.

Authored by Dr. David Kirk, Chief Scientist, NVIDIA

LARGE HEADER IMAGE NVIDIA's "Human Head" demo, created with the GeForce 8800 Ultra GPU, delivers a leap forward in realism, rendering the interplay of light and human skin at a level never before seen in real-time. Special thanks to actor Doug Jones for allowing the use of his likeness.

TRANSFORMATIONS // NVIDIA CORPORATION // WWW.NVIDIA.COM

Solving the World's Hardest Problems

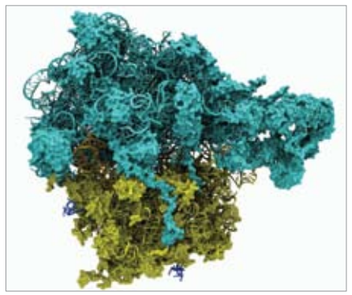
Since the introduction of NVIDIA Tesla, the company has seen a broad range of applications begin to harness the massive computational power of the GPU and use that to achieve unprecedented speed increases.

Somewhere deep down in the billions of cells that comprise the human body, something goes terribly wrong. A single cell starts to divide uncontrollably, launching a process that leads to cancer. Why does this happen? How can it be prevented? And once it's started, how can it be stopped? Thousands of scientists are pursuing these questions with the traditional tools of biological science: recombinant DNA, animal studies, and chemical experiments—all performed to divine the nature of cells and their internal machinery.

While traditional techniques observe the reactions of cells, some scientists have adopted a new approach that goes beyond observation to simulating and understanding what happens on the inside, at the most basic level of the cell. With new computing techniques and high-powered arrays of computers, scientists can simulate the basic processes of cells, from the atoms up, to understand how cells work and how they can go wrong. But the sheer scale of this task—simulating the parts of a cell, all the interactions of these parts, and how they respond to external factors—requires some of the most powerful computers in the world.

New computing techniques, like those being pioneered at the University of Illinois at Urbana-Champaign (UIUC), seek not only to expand the scope of simulation to more sophisticated situations, but also to provide a powerful new tool to any scientist in the pursuit of answers to how our bodies work.

UIUC is leading the way in using computers built with a GPU to put the equivalent of a large array of computers on the desk of any researcher. UIUC researchers have been using their NAMD (Nanoscale Molecular Dynamics) and VMD (Visual Molecular Dynamics) software running on NVIDIA GPUs to conduct simulations of nano-devices that can be used to sequence DNA in real-time and help reduce the cost of genomic medicine. They are seeing 100-240X speed increases and, more importantly, are now able to run these calculations at their desks, rather than queuing up to use large clusters in remote server rooms and waiting weeks for the results. (See www.ks.uiuc.edu/Research/vmd/)

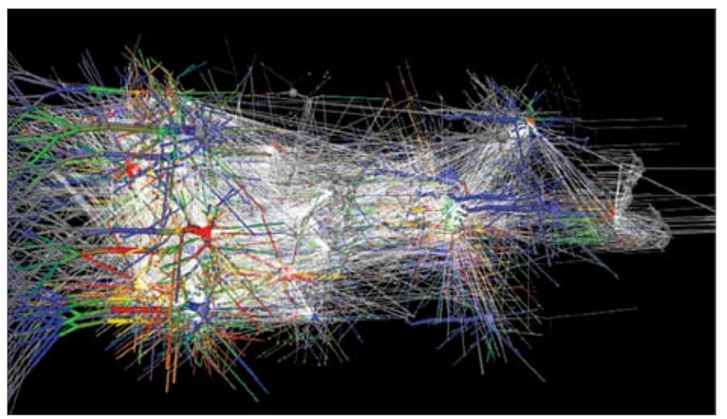


Similarly, Evolved Machines of Palo Alto, Calif. is using GPUs to reverse-engineer the circuits of the human brain, first to understand how they work and then to use those principles to create machines capable of similar functions (such as seeing and smelling). The company has, to date, achieved speed increases of more than 100X and has observed that a single desktop system containing two GPUs runs as fast as a 200-core cluster, at a fraction of the cost and power. (See www.evolvedmachines.com)

In the field of finance, Hanweck Associates has developed a real-time options implied-volatility engine, Volera. Using a single PC containing three GPUs, Volera can evaluate 150,000 options per second. With three Tesla S870 1U servers, Hanweck is able to calculate the entire US option market in real-time. (See http://www.hanweckassoc.com/)

These are just a few of the organizations using the massive parallel processing performance of the GPU to solve critical problems in fields like medicine, molecular biology, oil & gas exploration, and finance. Offering orders of magnitude improvements in simulation, NVIDIA's HPC solutions represent a discontinuity that will dramatically change the landscape of medicine and science.

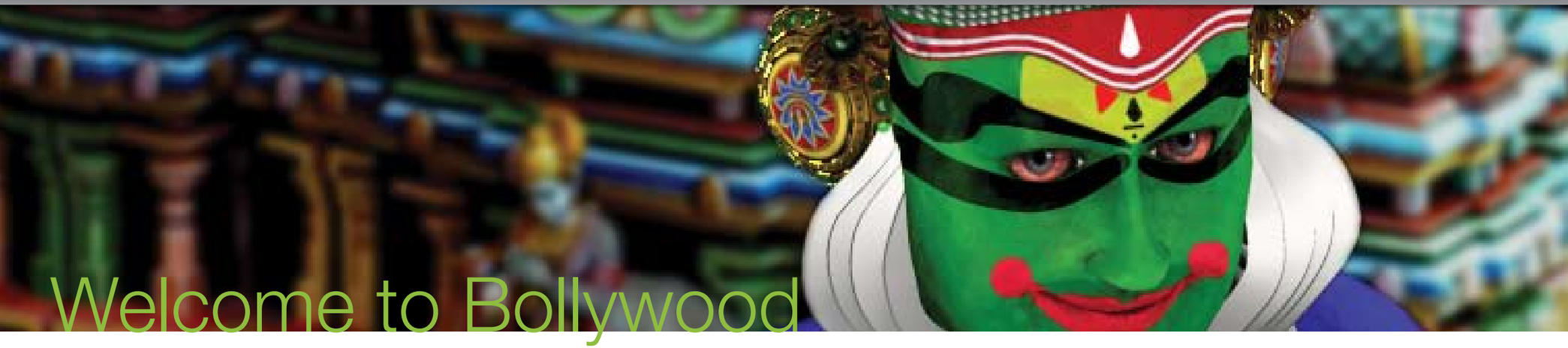
For more information, visit: www.nvidia.com/object/tesla_computing_solutions.html



FAR LEFT UIUC researchers are simulating complex molecular systems *(Image courtesy University of Illinois at Urbana-Champaign)*

THIS PAGE Evolved Machines is reverse-engineering brain circuits to develop a new paradigm for device technology *(Image courtesy Evolved Machines)*

TRANSFORMATIONS // NVIDIA CORPORATION // WWW.NVIDIA.COM



Welcome to Bollywood

From villages in the north known for dairy production to the idyllic beaches of the south, India is now witness to the booming industry of film and game production. Often compared to Los Angeles and its madcap film industry growth in the 1980s, these Indian markets are expanding as fast as studios can be built, artists trained, and movies distributed to the theaters.

Producing over 800 films annually, Mumbai is the hub of all this excitement, but production studios are springing up all over India. The warm climes of Goa are calling to many, as are Calcutta, Chennai, Hyderabad, and New Delhi. Combined with the very appealing tax incentives some Indian states are now offering these studios, growth is certain to continue.

Pune, already an established world hub for IT industries, has also set a more recent goal to become the animation and gaming hub of India and is well on its way with the establishment of the DSK School of Animation and Gaming, an offshoot of the world renowned Supinfocom animation school based in Valenciennes, France. Major players such as Disney Feature Animation and Sony Pictures Imageworks are already present and working closely with their Indian colleagues, producing content for worldwide consumption.

While the number of films produced is large, film production techniques in India had remained traditionally focused on 2D animation. Since the explosion of growth, we are now seeing 3D digital animation and visual effects capability taking root and starting to produce content for both domestic and global consumption. Groups ranging from the India-based National Association of Software and Service Companies (NASSCOM) and the Federation of Indian Chambers of Commerce and Industry (FICCI) to the Los Angeles-based Visual Effects Society and ACM Siggraph have all gotten onboard to help encourage this blossoming industry. It is no wonder why NVIDIA, with its significant share in the Digital Content Creation (DCC) market space worldwide recently launched an initiative called “Digital Bollywood.”

NVIDIA’s Digital Bollywood initiative is comprised of several key elements, including the sharing of expertise, training, and community development.

Expertise

Digital film production experience in India has been largely limited to a few individuals who cut their teeth on Hollywood and European productions and returned to India to start their own studios. Digital Bollywood is supplementing this expertise by bringing digital filmmaking expertise to India and exposing more Indian talent to the Hollywood experience. One of the ways NVIDIA is doing this is through the DCC Master’s Tour, a quarterly speaking tour by leaders in American and European digital content creation, having already hosted the likes of Electronic Arts and award-winning animator Andrew Daffy. The talks were attended by hundreds of young hopefuls and professionals alike in numerous cities. In turn, NVIDIA is also identifying promising students and young Indian filmmakers and sponsoring their attendance at western film festivals, where they can tap into expertise and build networks of contacts. NVIDIA is also deploying internal resources, like that of the NVIDIA Gelato development team, to train and consult with Indian studios in developing best practices for digital production.

Training

Indian culture places a premium on education. Digital Bollywood’s training and education efforts occur on two levels. First, NVIDIA has provided Indian film schools and academies with curricula and training materials on Gelato and other NVIDIA products. Second, NVIDIA has partnered with leading Indian film schools, like the DSK School of Animation and Gaming and the Whistling Woods Academy in Mumbai. The aim is to help develop a new generation of Indian filmmakers well-versed in digital technologies.

Community

Since NVIDIA does not compete with any of the Indian studios and since its products and technology are used by virtually all of them, the company is in a unique position to help foster symbiotic relationships between the studios. Working with business organizations like NASSCOM and FICCI, NVIDIA is active in sponsoring events and cooperation among the studios. NVIDIA has also been actively involved in supporting artistic and technical communities through involvement in ASIFA India, TASI, ABAI and other groups within the DCC communities.

A little more than a year into the effort, Digital Bollywood is beginning to bear fruit. NVIDIA Quadro has over 95% market share in India and the Gelato software renderer is gaining traction. The recently released film *Resident Evil: Extinction* (a Sony/Screen Gems release) which took in over $25 million at the box office during its opening weekend in the U.S., contained visual effects rendered with Quadro and Gelato by Mumbai-based Anibrain.

A key to the success of the Digital Bollywood initiative is to take the long-term view. From the start, NVIDIA realized that occasionally bringing in executives from North America would not suffice. Follow-through and long-term relationships are key to building the market. By dedicating people knowledgeable about the film industry to the project and leveraging substantial technical capabilities resident in India, NVIDIA is well-positioned to develop and sustain the effort.

Laura Dohrmann of NVIDIA contributed to this article.

Bollywood Facts

- The 2006 release of summer blockbuster “Krrish” was the first film produced where all the visual effects were done by India-based film studios.
- Celebrated Hindi actor Shah Rukh Khan has appeared in over 60 films since 1992. His product endorsements are some of the most valued in India today.
- Bollywood film trailers and music videos are the most popular mobile phone downloads after sports and fashion in India.
- The “Indian Oscars” are the *Filmfare Awards*, sponsored by *Filmfare* magazine, first given out in 1954.
- The first Indian film was *Raja Harishchandra*, a silent movie made in 1913.

LARGE HEADER IMAGE Image created by Deepak Hargaonkar of Indore, the first prize winner in the NVIDIA Gelato Imaging Contest. **SMALL HEADER IMAGES** Mumbai-based Anibrain coupled NVIDIA Gelato film rendering software with NVIDIA Quadro GPUs to create realistic visual effects for *Resident Evil: Extinction* (a Sony/Screen Gems release). Images courtesy Anibrain.

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**

NVIDIA