MOOR
INSIGHTS & STRATEGY

NVIDIA's 2nd Generation Maximus

Dawn of the "Hybrid Designer"

Patrick Moorhead, President & Principal Analyst, Moor Insights & Strategy
November, 2012

This paper was commissioned by NVIDIA.

EXECUTIVE SUMMARY

All of today's sophisticated physical products like cars and media content like movies are designed then simulated on high-performance PCs, called workstations. Today, the design and simulation process is dictated by the technology, not the natural or most efficient way to work. Industry professionals in entertainment, media, sciences, and industrial design fields have been hindered by numerous technological restraints that limit their full potential. Current processes are so inefficient that two separate steps and sometimes corporate divisions still exist, split between *design* and *simulation*.

Additionally, the massive workloads required for multilayer rendering and heavy-duty computing with huge datasets claim valuable time from engineers who act on tight deadlines, resulting in overloaded system resources, long wait times and bottlenecks on shared machines, labor and cost inefficiencies, and maybe most importantly, delayed time to market.

With the first generation Maximus, NVIDIA introduced a breakthrough in hybrid, parallel computing on a two card GPU configuration that work in concert for visualization and computation. Visualization is accomplished on an NVIDIA Quadro card and computation is done on an NVIDIA Tesla card inside the workstation. Design and ray trace rendering, a heavy computational technique used to simulate light interactions in 3D environments, were performed simultaneously in real time to create incredibly rich, realistic graphics, enabling the *user* to maintain its intended functionality as the driver of operations. The CPU reduction from GPU loads under Maximus enabled multitasking freedom to users who are no longer held hostage by their workstations and were able to fully engage all of their creative skills and resources.

The recently-announced 2nd generation Maximus significantly increases performance and more importantly performance per watt and can fundamentally change the design and simulation workflow. Compared to prior Fermi-based architecture designs, the new Kepler-based cards feature 3X performance per watt, simplified programming features, twice the graphics memory to further decrease CPU loads, reference to more textures, and film-style methods that soften hard, jagged edges for incredibly realistic virtual images.

Additionally, NVIDIA's strategic partnerships with application developers are expanding the number of software solutions and are increasingly easing programming on the GPU, unleashing its full capacity for graphic visualizations and conceptualizations with computing speeds unmatched in the industry.

With this substantial increase in performance per watt, the split between design, simulation, and rendering starts to become obsolete, considered "old school." This capability will usher in the age of the “hybrid designer”, who, within the same app will be able to design *and* simulate in real-time. This speeds up TTM (time to market) even further than the first version of Maximus provided and is very valuable to design and development organizations. As OEMs partners like HP, Dell and Lenovo develop even more workstation solutions, more mainstream users could have access to high performance computing that is configurable to need and cost. This could even drive an increase in the size of the marketplace for NVIDIA and its Maximus ecosystem.

NVIDIA’S MAXIMUS VISION

Released in 2011, NVIDIA’s first generation Maximus technology introduced the future of combined visualization and computing by the GPU, and its upcoming second generation architecture will further revolutionize design and engineering workstations and workflows. NVIDIA envisions a complete rebirth of the traditional workstation, bringing supercomputing compute performance and visuals by parallel, graphics processing to the business workplace, university and ultimately the prosumer. Second generation Maximus will impact professionals across a wide variety of industries, but manufacturing design, media and entertainment, and energy will see immediate benefits from currently available applications. In the future, any advanced application that has heavy visualization and computation could take advantage of the 2nd generation Maximus.

Manufacturing Industries[1]

What would it be worth to a durable goods company to reduce product TTM measured by months, combine design and simulation groups, and enhance the visual identity? With 2nd generation Maximus, developers will see shortened cycles, reduced delays to market, and a highly interactive creative process with minimal rendering time. Engineering analysis on the designers desktop will improve product quality through more design variations and simulations while reducing cost and bottlenecks in system usage. Manufacturing designers are empowered to think and create beyond limits. When testing new automobile or aircraft concepts with Maximus, endless quantities of components and parts are available for interactive analysis with real world environments, and rapid ray trace rendering and scaling generates immediate feedback for reengineering the model or moving forward in production.

[1] http://www.NVIDIA.com/object/maximus-for-manufacturing-industries.html

[2]Media and Entertainment Industries

Second generation Maximus will produce faster and higher quality productions while decreasing costs. Designers can do transform editing, layering, rendering, and video encoding, and do this in *real-time*. They can add more 3D effects and depth to dramatic visual content while reducing production costs and time. Character and scene features in animation and film are responsive to elements like rain and wind, which are produced in elaborately detailed layers that can be *real-time* edited and previewed quickly in numerous situations.

[3]Energy Industries

2nd generation Maximus will increase competitive advantages with computational speeds for accurate interpretation of complex data sets and high resolution geological surveys. Quadro graphics enable greater visualization and clarity of geological formations which are critical in mitigating risks and managing costs in oil and gas exploration. Tesla provides the computational horsepower for these workloads.

Similar advantages can be found through second generation Maximus optimized workstations for medical imaging and other industries that require confident, timely analysis and action based on visual data.

As the full efficiency and creative benefits from second generation Maximus are realized, the software and hardware ecosystem will start growing exponentially. As the hardware and software ecosystem grows, “Maximus”-certified workstations and applications will become a required element for all commercial bids and workflows.

TODAY’S CHALLENGES

[4],[5]Despite the endless possibilities of graphics processing, great challenges accompany its rapid evolution. For design engineers in manufacturing who run operations in CAD

[2] http://www.NVIDIA.com/object/maximus-media-entertainment.html

[3] http://www.NVIDIA.com/object/maximus-for-energy-seismic-industry.html

[4] http://blogs.NVIDIA.com/2011/11/taking-the-suspense-and-waiting-out-of-simulation/

[5] http://www.deskeng.com/virtual_desktop/?p=4604

modeling and manage incredibly large sets of data to create complex simulations, an immense investment is required for computation and system processing. Engineers are forced to strategically allocate time to run these jobs, often at night or during lunch breaks to avoid machine overloads. In addition to slowing down equipment, team productivity is severely affected as these engineers could be performing other tasks.

In environments where engineers have access to high-performance machines, CPU intensive jobs lay in bottlenecked queues on shared resources, and critical decision making and time sensitive adjustments or corrections must wait on full completion of processing tasks, often resulting in wasted efforts. Content creation professionals in media and entertainment as well as those employed in the energy industry face similar obstacles that all result in delayed time to market, labor inefficiencies, and competitive imbalances in lost opportunities.

Workstation designs are uneconomical as well, even though GPUs turbo charge application performance and revolutionize graphics potential. In a traditional workstation, one CPU handles design and engineering workflows while a second CPU is often used to back up the first by handling multi-threaded rendering and simulation processes. In attempts to multitask, serial workflows and single and dual socket designs require waiting by the GPU as the CPU determines the order of tasks to be handled by the appropriate chip.

While NVIDIA’s CUDA provides a comprehensive GPU development environment, programming for the first generation Maximus remains challenging and limited to the realm of the most sophisticated “ninja programmers”. Unlike a CPU programmer, today’s Maximus application programmer needs to comprehend two different kinds of memory, system and GPU memory, which is a challenge.

While first generation Maximus provided tremendous benefit to designers and engineers, it could still be improved by providing even greater performance per watt and simplifying the programming model. This in turn, could improve TTM and creativity, significantly improving the bottom line results. This is where 2nd generation Maximus comes into play.

APPROACH & APPLICATION IMPACT

2nd generation Maximus will help mitigate many of the challenges of users and machines relying on graphics processing while revolutionizing creative design and engineering. To achieve the speeds, efficiencies, and design possibilities promised under the 2nd generation Maximus architecture, NVIDIA will utilize its most sophisticated GPU architecture, Kepler. NVIDIA pioneered computing on the GPU, and by many measurements, NVIDIA has now created the world's fastest, most efficient cards on the market.

[6]Set for release in Q4 2012, the Kepler-powered Tesla K20 compute card features 2,496 computing cores and reduced control logic through its SMX (streaming multiprocessor) design, which equates to 3X performance per watt over Tesla's prior Fermi architecture and further expands its hybrid computing potential on a wider scope of applications. Dynamic Parallelism and Hyper-Q are features that simplify and advance parallel programming, decrease idle times, and increase GPU utilization. GPU threads automatically spawn to lessen reliance on CPU resources, multiple CPU cores access CUDA cores on a single Kepler GPU, and greater percentages of space are applied to processing cores.

[7]The Kepler-powered Quadro K5000 visualization card creates industry leading performance and efficiency in graphics visualization. Design engineers will benefit from the increased operations and large memory capacity to interact and analyze during the design of complex models. Bindless Textures is a feature that reduces CPU overhead, increases textures (over 1 million) available to shaders directly in graphics memory, and offers higher image quality and more realistic texture detail in scenes.

The following are representative benefits according to NVIDIA's testing of a brief sample of available solutions that are reshaping and revolutionizing workloads.

[8]Manufacturing Applications:

Photorealistic Rendering and Design	Engineering Analysis and Design	Ray Tracing
3DS Max 2012	Ansys	Catia Live Rendering
9x rendering performance increase	5x performance	650% speed increase
Quadro 6000/Tesla C2075 vs. 8 CPU cores	8 CPU cores/Tesla C2075 vs. 2 CPU cores	Quadro 6000/Tesla C2075 vs. 8 CPU cores

[6] http://www.NVIDIA.com/content/tesla/pdf/nv-ds-teslak-family-jul2012-lr.pdf

[7] http://www.NVIDIA.com/object/quadro-k5000.html

[8] http://www.NVIDIA.com/object/maximus-for-manufacturing-industries.html

[9]Media and Entertainment:

Design	Ray Traced 3D Rendering
[10]Adobe Premier Pro CS5.5.2 with Adobe Mercury Playback Engine (MPE)	Adobe After Effects CS6
900% performance increase	2,500% relative increase
Quadro 2000/Tesla C2075 vs. 8 CPU cores	Quadro 6000/Tesla C2075 vs. 12 CPU cores

CASE STUDY

NVIDIA and its application development partners have offered multiple demos and reactions to Maximus enabled technology, all equating to incredible graphic results in less time carried out by systems and users acting at their highest capacity. One particular case study highlighted the removal of creative barriers in auto design and testing.

Mercedes Benz

[11]Annually at the L.A. Auto Show in Los Angeles, concept car designers meet to showcase the future in art and function of automotive possibility. In 2012, Mercedes entered the L.A. Auto Show Design Challenge taking advantage of NVIDIA's Maximus architecture. In the past, Mercedes' design teams were restricted by the capability of their equipment, often scrapping ideas due to the hundreds of hours of intensive frame rendering required by the digital reproduction of real life lighting elements. With Maximus, Quadro and Tesla GPUs ran Autodesk 3ds Max, Bunkspeed's Move/Drive with NVIDIA iray®, Maya, Alias, Photoshop, and Adobe After Effects software for the development of the Silver Arrow. Workstations simultaneously performing complex visualizations broke the tedious serial design and simulation process and offered teams to design scene by scene with rapid feedback and multiple outputs.

According to Alan Barrington, the Mercedes-Benz Advanced Design Center California designer responsible for creating the car's animation commented that his team no longer has to settle or compromise in creativity. "With the parallel processing capabilities enabled by the NVIDIA Maximus systems, we can now be *10 times more creative*...In a given amount of time; we can explore so many different options and get to a better end product."

[9] http://www.NVIDIA.com/object/maximus-media-entertainment.html

[10] http://www.NVIDIA.com/docs/IO/40049/WP-AdobeandCUDA.pdf

[11] http://www.NVIDIA.com/content/quadro/maximus/case-study/NVIDIA-Maximus-success-story-Mercedes-Benz-FINAL.pdf

By parallelizing design and simulation, Mercedes was able to dramatically enhance creativity, which could ultimately provide them with improved competitiveness through decreased TTM and increased consumer demand and pricing power through a more desirable design. While the Mercedes example is a good one, it applies to other manufacturing, digital media, and energy industries.

As seen in the Mercedes Benz example, Maximus provided them measured value. There are more compelling examples that NVIDIA has on display on their website. One of these is Daniel Simon, an auto design firm who reduced render times on a movie project from 15 minutes to 5 per frame or a reduction of 2/3rds. This ultimately saved 11 days of production time, which, again, results in real business impact.

ECOSYSTEM IMPLICATIONS

The implications to designers and developers from second generation Maximus have already been explored in this paper, but what about the rest of the ecosystem?

There is already an installed base of design and simulation hardware and software and this will need to be replaced. As the market becomes more accessible through lower cost, the market footprint will grow, increasing the size of the total available market (TAM) for everyone. Maximus also challenges where the compute is performed, particularly with the CPU versus the GPU.

The workstation hardware and ecosystem will obviously benefit from successful 2nd generation Maximus implementations. In hardware, this primarily includes workstation OEMs HP, Dell and Lenovo. These OEMs will be able to go to their installed workstation base and offer upgrades from older, less sophisticated legacy systems and Maximus configurations to the 2nd generation Maximus.

Every one of these new installations will include new application software, too, increasing licensing revenue and profits. As most design and simulate packages come from different companies, there is a new ISV opportunity and threat. Some ISVs will expand their competitive footprint by adding design to their simulation package or simulation to their design package.

2nd generation Maximus also significantly destabilizes the harmony between CPU and GPU. Many traditional workstations perform visualization on the GPU and simulation on the CPU. If users can attain superior ROI by doing most of the compute on the GPU, what does this mean for the CPU? Users can opt for lower end CPUs for an optimized workstation for their workflows. Intel is already reacting with their new Xeon Phi product line with MIC (Many Integrated Cores), but for now, NVIDIA has a head start.

NVIDIA can obviously benefit from a successful rollout. If users choose to do the heavy compute on Maximus versus the CPU, NVIDIA could double their “basket” size per workstation. Workstations routinely have $1,000-2,000 CPUs inside them today, and many of them have two and even four sockets.

FINAL THOUGHTS

Historically, major technological inflection points only become important as they provide the users or buyers with significant improvements that motivate changes in the status quo. Ecosystems must also see the benefit to the change for the technology to be fully realized. When deployed on the right workloads with the right application, 2nd generation Maximus can be one of those game changers. Maximus can improve the way companies design *and* develop products which in turn could change the competitive dynamics in manufacturing, media and entertainment, and energy industries. As the economics improve to roll out this capability to an even larger base, Maximus even has the capability to radically impact society as a whole by bringing much better end products to market quicker and more economically.

Author

Patrick Moorhead, Principal Analyst, Moor Insights & Strategy

Inquiries

Please contact us at the email address above if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Licensing

Disclosures

This paper was commissioned by NVIDIA.

DISCLAIMER

The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.