Whitepaper

NVIDIA GeForce GTX 750 Ti

Featuring First-Generation Maxwell GPU Technology, Designed for Extreme Performance per Watt

V1.1

Table of Contents

Introduction

Whether it's the amazing special effects in the latest Hollywood blockbuster movie, near-photorealistic 3D game environments with lifelike characters, or a media-rich website with higher resolution images and video, consumer demand for more stunning graphics continues to increase with every passing year. To meet this challenge, NVIDIA's graphics processors have continued to evolve, with each generation incorporating new features and becoming more powerful. Our Kepler GPU architecture was introduced in early 2012, delivering groundbreaking performance and power efficiency. Kepler GPUs powered the world's fastest gaming PCs and workstations, as well as supercomputers and cloud gaming servers. Kepler GPU architecture was also implemented in the Tegra K1 system-on-a-chip family to enable industry-leading visual computing capabilities in smartphones, tablets and even the infotainment systems found in cars.

In order to take graphics to the next level of visual realism however, NVIDIA engineers recognized early on that we had to make our next architecture even more efficient than Kepler.

NVIDIA's first-generation "Maxwell" architecture implements a number of architectural enhancements designed to extract even more performance per watt consumed. The first Maxwell-based GPU is codenamed "GM107" and designed for use in power-limited environments like notebooks and small form factor (SFF) PCs. These SFF systems are often used for gaming and home entertainment, with the most recent example being Valve Software's recently announced Steam Machines initiative. The first graphics card that is based on the GM107 GPU is the GeForce GTX 750 Ti. Because of GM107's remarkable architectural efficiency, at 1080p resolution a GeForce GTX 750 Ti will frequently match the performance of our flagship GPU from four years ago, the GeForce GTX 480, but with only a 60W TDP, consumes a fourth of the power.

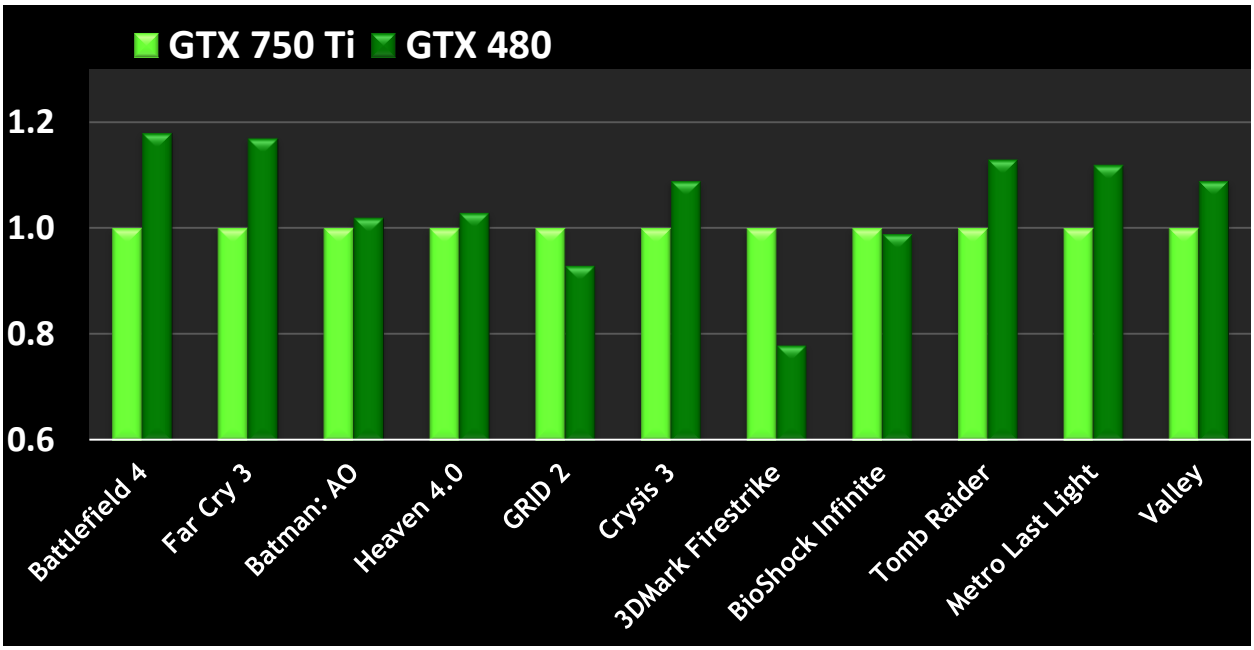

Figure 1: *GeForce GTX 750 Ti performs evenly with GTX 480 in many of today's top titles*

GM107 is the first GPU built using the first-generation Maxwell architecture. These first-generation Maxwell products are focused on low power operation. We'll also introduce higher performing second-generation Maxwell GPUs addressing the performance and enthusiast graphics segments at a later date.

The Soul of Maxwell: Improving Performance per Watt

During the course of transitioning Kepler from a GPU used in PCs, workstations, and supercomputers down to a mobile chip that can fit in your pocket, we learned a lot about not only reducing GPU power consumption overall, but also how we can extract more performance from our architecture at the same power level. Everything we learned through this effort went into Maxwell.

Maxwell introduces an all-new design for the Streaming Multiprocessor (SM) that dramatically improves performance per watt and performance per area. Although the Kepler SMX design was extremely efficient for its generation, through its development NVIDIA's GPU architects saw an opportunity for another big leap forward in architectural efficiency; the Maxwell SM is the realization of that vision. Improvements to control logic partitioning, workload balancing, clock-gating granularity, scheduling, number of instructions issued per clock cycle, and many other enhancements allow the Maxwell SM (also called "SMM") to far exceed Kepler SMX efficiency. The new Maxwell SM architecture enabled us to increase the number of SMs to five in GM107, compared to two in GK107, with only a 25% increase in die area. We'll discuss the changes made to the Maxwell SM in greater detail in the "Next-Generation Maxwell SM" section later in this document.

Maxwell also boasts a dramatically larger L2 cache design; 2048KB in GM107 versus 256KB in GK107. With more cache located on-chip, fewer requests to the graphics card DRAM are needed, thus reducing overall board power and improving performance.

In addition to the changes above, NVIDIA engineers aggressively tuned the implementation of each unit in the Maxwell GPU down to the transistor level, to maximize energy efficiency.

The end result of all of these efforts is that Maxwell is able to deliver 2 times the performance/watt of Kepler, using the same 28nm manufacturing process.

GM107 Maxwell Architecture In-Depth

From a graphics features perspective, our first-generation Maxwell GPUs offer the same API functionality as Kepler GPUs. At the high level, Maxwell also implements multiple SM units within a GPC (Graphics Processing Cluster), and each SM includes a Polymorph Engine and Texture Units, while each GPC includes a Raster Engine. ROPs are still aligned with L2 cache slices and Memory Controllers. Internally, all the units and crossbar structures have been redesigned, data flows optimized, power management significantly improved, and so on.

The GM107 GPU contains one GPC, five Maxwell Streaming Multiprocessors (SMM), and two 64-bit memory controllers (128-bit total). This is the full implementation of the chip, and is the same configuration we ship with the GeForce GTX 750 Ti.

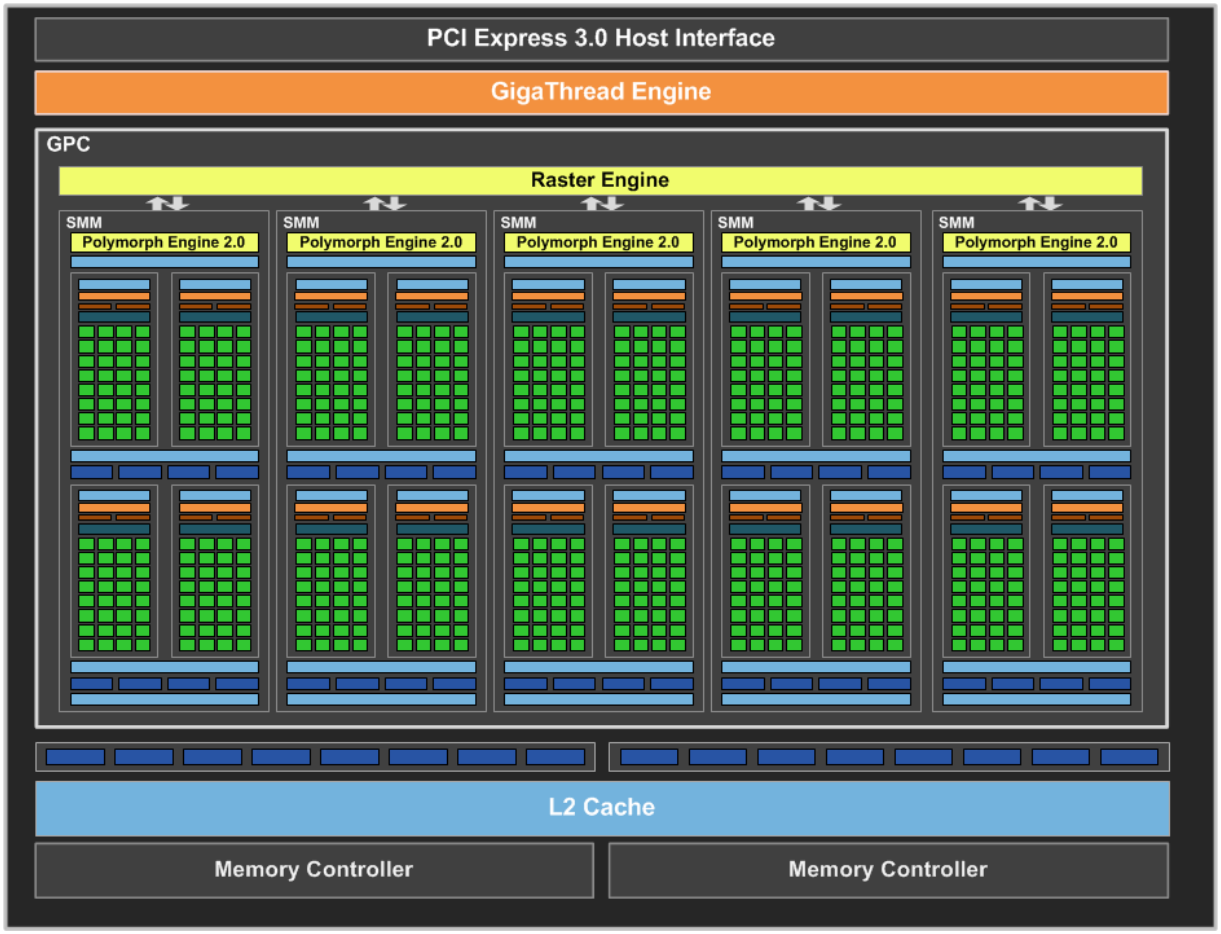

Figure 2: *GM107 Full-Chip Block Diagram*

The following table provides a high-level comparison of Maxwell vs. our prior generation GK107 Kepler GPU:

GPU	GK107 (Kepler)	GM107 (Maxwell)
CUDA Cores	384	640
Base Clock	1058 MHz	1020 MHz
GPU Boost Clock	N/A	1085 MHz
GFLOPs	812.5	1305.6
Texture Units	32	40
Texel fill-rate	33.9 Gigatexels/sec	40.8 Gigatexels/sec
Memory Clock	5000 MHz	5400 MHz
Memory Bandwidth	80 GB/sec	86.4 GB/sec
ROPs	16	16
L2 Cache Size	256KB	2048KB
TDP	64W	60W
Transistors	1.3 Billion	1.87 Billion
Die Size	118 mm²	148 mm²
Manufacturing Process	28-nm	28-nm

We've discussed the basic aspects of our architecture – such as the dataflow from the host PCI Express interface through the GigaThread engine, basic operation of Polymorph and Raster units, etc. - in greater detail in our Kepler and Fermi whitepapers. In case you need additional background information on how they operate, we highly recommend reading those documents first. We'll be providing more detail on the changes introduced in SMM on the following pages.

Next-Generation Maxwell SM

The primary contributor to Maxwell's improved efficiency is the new Maxwell SM architecture, SMM. This new SM architecture achieves much higher power efficiency and delivers 35% more performance per CUDA Core on shader-limited workloads. Achieving these results required a number of major changes to the architecture. The SM scheduler architecture and algorithms have been rewritten to be more intelligent and avoid unnecessary stalls, while further reducing the energy per instruction required for scheduling.

The organization of the SM has also changed. Each SM is now partitioned into four separate processing blocks, each with its own instruction buffer, scheduler and 32 CUDA cores. The Kepler approach of having a non-power-of-two number of CUDA cores, with some that are shared, has been eliminated. This partitioning simplifies the design and scheduling logic, saving area and power, and reduces computation latency.

Pairs of processing blocks share four texture filtering units and a texture cache. The compute L1 cache function has now also been combined with the texture cache, and shared memory is a separate unit (similar to the approach used on G80, the first CUDA capable GPU), that is shared across all four blocks.

Overall, with this new design, each “SM” is significantly smaller while delivering about 90% of the performance of a Kepler SM, and the smaller area enables us to implement many more SMs per GPU. Comparing GK107 versus GM107 total SM related metrics, GM107 has five versus two SMs, 25% more peak texture performance, 1.7 times more CUDA cores, and about 2.3 times more delivered shader performance.

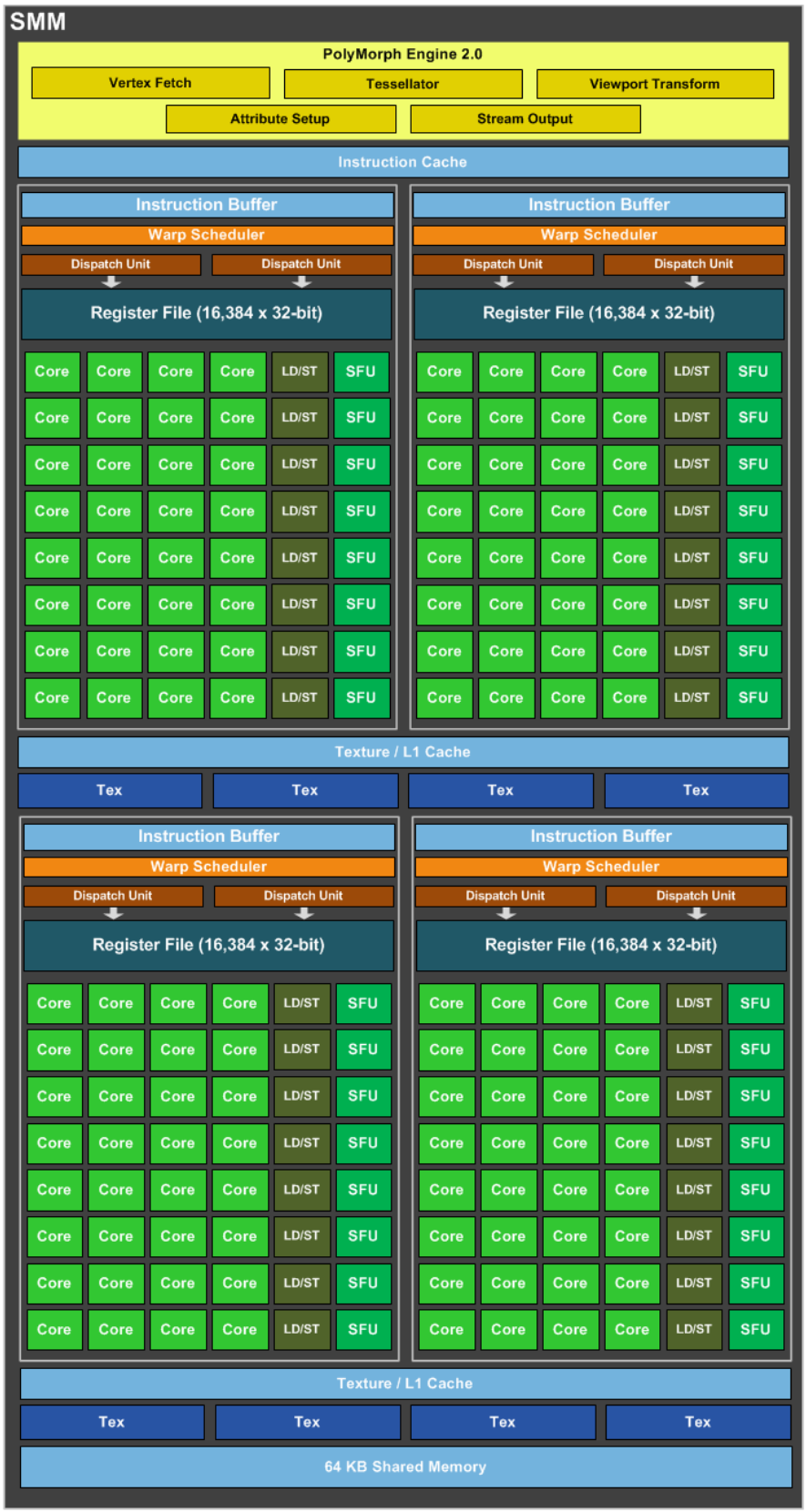

Figure 3: *Maxwell SM Block Diagram*

Memory System

For GM107, to achieve its goal of significantly higher performance with the same memory width as GK107, it was also important to invest in memory system enhancements. On-chip memory system bandwidth was increased along with improvements in efficiency of the design. In addition, the large 2MB L2 cache configuration (larger than any previous GPU design) is highly effective at reducing memory bandwidth demand and ensuring that DRAM bandwidth is not a bottleneck.

New Video Capabilities

One of Kepler's key innovations over prior GeForce GPUs was its hardware-based H.264 video encoder, NVENC. By integrating dedicated hardware circuitry for video encoding/decoding (rather than using the GeForce GPU's CUDA Cores) NVENC provided a dramatic performance speedup for H.264 encoding while consuming less power.

We leveraged Kepler's NVENC encoder to introduce ShadowPlay to GeForce GTX 600 series and GTX 700 gamers last fall, allowing them to record their favorite gaming moments for anyone to see. Since launching ShadowPlay, over 3 million videos have been captured, with gamers posting them to YouTube or even streaming their gameplay footage live over Twitch.

To improve video performance, Maxwell features an improved NVENC block that provides faster encode (6-8X real-time for H.264 vs. 4x real-time for Kepler) and 8-10X faster decode, and thanks to a new local decoder cache, higher memory efficiency per stream for video decoding, resulting in lower power for video decode.

Maxwell also features a new GC5 power state that's been tailored to reduce the GPU's power consumption specifically for light workload cases like video playback. GC5 is a low power sleep state that provides considerable power savings over prior GPUs for these scenarios.

Conclusion

Given the increased challenges in developing ever smaller semiconductor manufacturing process nodes, NVIDIA engineers recognized early on that in order for PC graphics to continue to evolve, our Maxwell architecture would have to become more efficient. Simply building a bigger Kepler wouldn't be enough. Instead we set out to deliver groundbreaking performance per watt with Maxwell.

In order to improve performance while minimizing wasted power, we've grouped the SM into quads, each with its own dedicated resources for scheduling and instruction dispatch. We've also dramatically increased the size of the L2 cache, providing an additional storage buffer that is shared across the GPU for texture requests, atomic operations or anything else, saving trips to memory.

With the changes made in Maxwell's new SMM, the GPU's hardware units are utilized more often, resulting in greater performance and power efficiency. The GeForce GTX 750 Ti delivers over 1.7X more performance than GK107, and with a TDP of just 60W!

Twenty years ago, PCs were largely confined to a 2' tower that resided in the home office. Today, PCs can be found in any room throughout the home and can be as small as a shoe box. Up until now, gaming on one of these tiny PCs has been a rather mundane experience, as most users in this segment typically settle for a CPU with integrated graphics, resulting in low frame rates and ultra-low graphics settings. However, thanks to Maxwell's tremendous power efficiency, gamers with home theater and other small form factor PCs no longer have to compromise to get a good gaming experience at 1080p.

This means that you can easily plug the GeForce GTX 750 Ti into a wide range of PCs – no need to worry about upgrading your power supply in most cases. As a result, you can take a basic, off-the-shelf PC and transform it into a competent gaming PC. And because the GeForce GTX 750 Ti consumes so little power, it runs extremely quiet and generates very little heat, making it perfect for use in a home theater PC. It's the world's fastest graphics card that doesn't require a power connector.

Because of our intense focus on performance/watt, Maxwell is the world's most efficient GPU. As a result, gamers can enjoy their favorite games in almost any form factor like never before.