NVIDIA® KEPLER GK110 NEXT-GENERATION CUDA® COMPUTE ARCHITECTURE

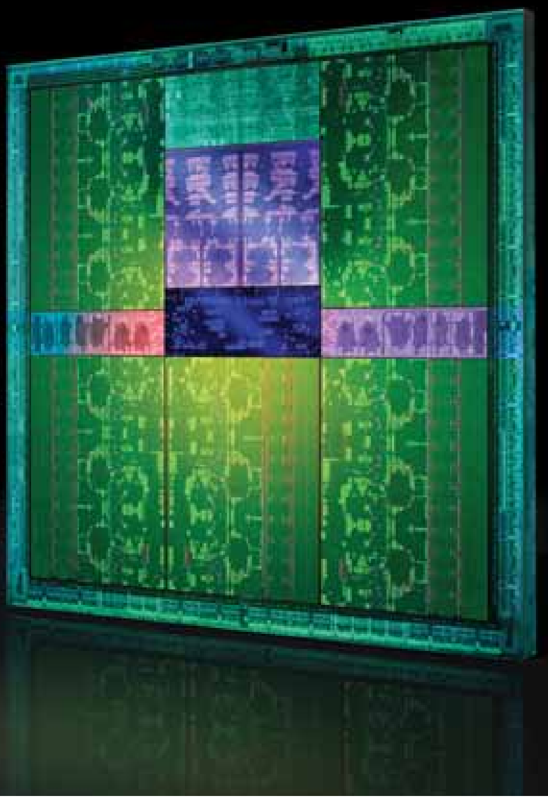

Figure 1: Kepler GK110 GPU- World's fastest and most power efficient x86 accelerator

FASTEST, MOST EFFICIENT HPC ARCHITECTURE

With the launch of Fermi GPU in 2009, NVIDIA ushered in a new era in the high performance computing (HPC) industry based on a hybrid computing model where CPUs and GPUs work together to solve computationally-intensive workloads. And in just a couple of years, NVIDIA Fermi GPUs powers some of the fastest supercomputers in the world as well as tens of thousands of research clusters globally. Now, with the new Kepler GK110 GPU, NVIDIA raises the bar for the HPC industry, yet again.

Comprised of 7.1 billion transistors, the Kepler GK110 GPU is an engineering marvel created to address the most daunting challenges in HPC. Kepler is designed from the ground up to maximize computational performance with superior power efficiency. The architecture has innovations that make hybrid computing dramatically easier, applicable to a broader set of applications, and more accessible.

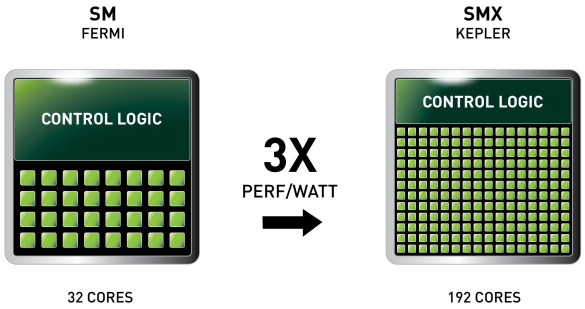

Figure 2: SMX: 192 CUDA cores, 32 Special Function Units (SFU), and 32 Load/Store units (LD/ST)

Kepler GK110 GPU is a computational workhorse with teraflops of integer, single precision, and double precision performance and the highest memory bandwidth. The first GK110 based product will be the Tesla K20 GPU computing accelerator.

This technical brief is designed to quickly summarize three of the most important features in the Kepler GK110 GPU: SMX, Dynamic Parallelism, and Hyper-Q. For further details on additional architectural features, please refer to the Kepler GK110 Whitepaper.

SMX - NEXT GENERATION STREAMING MULTIPROCESSOR

At the heart of the Kepler GK110 GPU is the new SMX unit, which comprises of several architectural innovations that make it not only the most powerful Streaming Multiprocessor (SM) we've ever built but also the most programmable and power-efficient.

DYNAMIC PARALLELISM — CREATING WORK ON-THE-FLY

One of the overarching goals in designing the Kepler GK110 architecture was to make it easier for developers

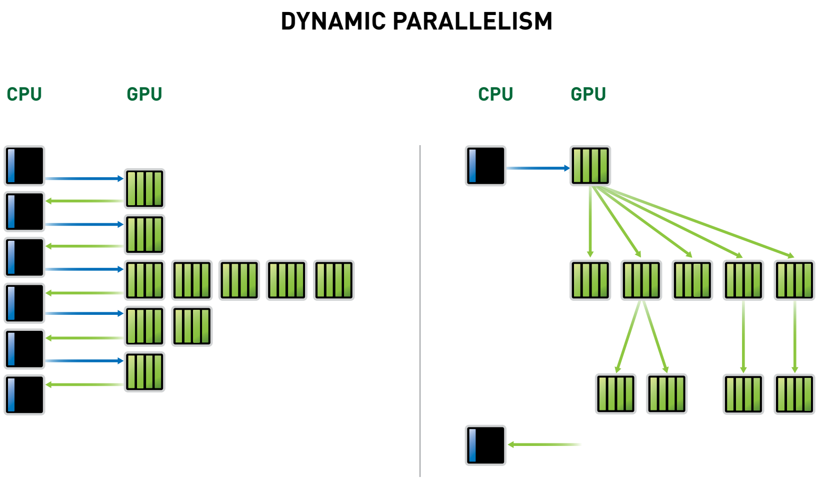

Figure 3: Without Dynamic Parallelism, the CPU launches every kernel onto the GPU. With the new feature, Kepler GK110 GPU can now launch nested kernels, eliminating the need to communicate with the CPU.

to more easily take advantage of the immense parallel processing capability of the GPU.

To this end, the new Dynamic Parallelism feature enables the Kepler GK110 GPU to dynamically spawn new threads by adapting to the data without going back to the host CPU. This effectively allows more of a program to be run directly on the GPU, as kernels now have the ability to independently launch additional workloads as needed.

Any kernel can launch another kernel and can create the necessary streams, events, and dependencies needed to process additional work without the need for host CPU interaction. This simplified programming model is easier to create, optimize, and maintain. It also creates a programmer friendly environment by maintaining the same syntax for GPU launched workloads as traditional CPU kernel launches.

Dynamic Parallelism broadens what applications can now accomplish with GPUs in various disciplines. Applications can launch small and medium sized parallel workloads dynamically where it was too expensive to do so previously.

HYPER-Q — MAXIMIZING THE GPU RESOURCES

Hyper-Q enables multiple CPU cores to launch work on a single GPU simultaneously, thereby dramatically increasing GPU utilization and slashing CPU idle times. This feature increases the total number of connections between the host and the the Kepler GK110 GPU by allowing 32 simultaneous, hardware managed connections, compared to the single connection available with Fermi. Hyper-Q is a flexible solution that allows connections for both CUDA streams and Message Passing Interface (MPI) processes, or even threads from within a process. Existing applications that were previously limited by false dependencies can see up to a 32x performance increase without changing any existing code.

Hyper-Q offers significant benefits for use in MPI-based parallel computer systems. Legacy MPI-based algorithms were often created to run on multi-core CPU-based systems. Because the workload that could be efficiently handled by CPU-based systems is generally smaller than that available using GPUs, the amount of work passed in each MPI process is generally insufficient to fully occupy the GPU processor.

While it has always been possible to issue multiple MPI processes to concurrently run on the GPU, these processes could become bottlenecked by false dependencies, forcing the GPU to operate below peak efficiency. Hyper-Q removes false dependency bottlenecks and dramatically increases speed at which MPI processes can be moved from the system CPU(s) to the GPU for processing.

Hyper-Q promises to be a performance boost for MPI applications.

CONCLUSION

Kepler GK110 GPU is engineered to deliver ground-breaking performance with superior power efficiency while making GPUs easier than ever to use. SMX, Dynamic Parallelism, and Hyper-Q are three important innovations in the Kepler GK110 GPU to bring these benefits to reality for our customers. For further details on additional architectural features, please refer to the Kepler GK110 Whitepaper at **http://www.nvidia.com/object/nvidia-kepler.html**.

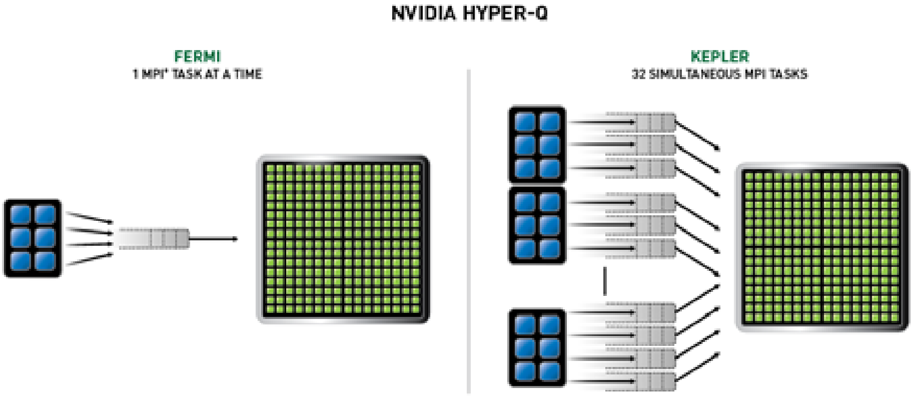

Figure 4: Hyper-Q allows all streams to run concurrently using a separate work queue. In the Fermi model, concurrency was limited due to intra-stream dependencies caused by the single hardware work queue.

To learn more about NVIDIA Tesla, go to **www.nvidia.com/tesla**