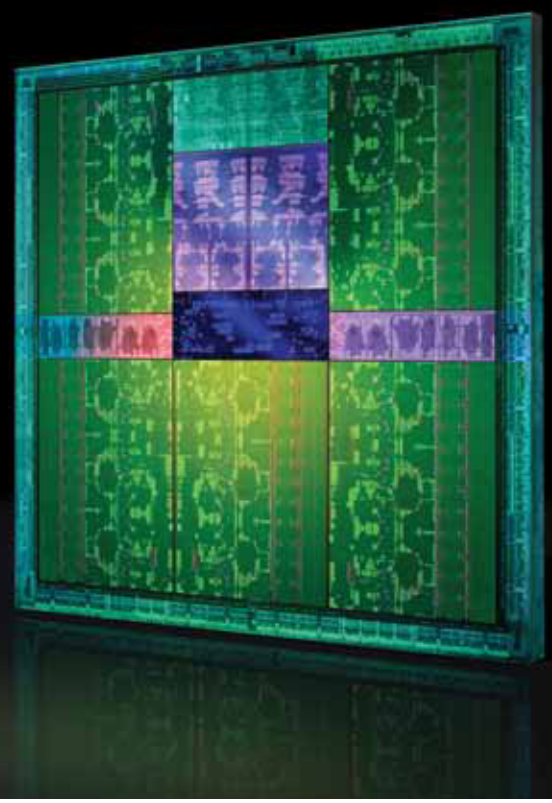

英伟达™ (NVIDIA®) KEPLER GK110 下一代 CUDA® 计算架构

最快、效率最高的 HPC 架构

随着 Fermi GPU 在 2009 年的推出，英伟达在高性能计算（HPC）行业迎来了一个新的时代，其基于混合计算模型，其中 CPU 和 GPU 协同工作来解决计算密集型工作负载。在短短几年，英伟达™ Fermi GPU 增强了世界上一些速度最快的超级计算机以及全球数以万计的研究集群。目前，有了新 Kepler GK110 GPU，NVIDIA 又一次提高了 HPC 行业的标准。

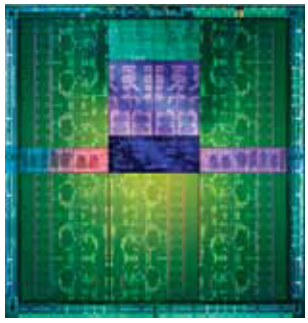

图 1：Kepler GK110 GPU- 世界上速度最快和最省电的 x86 加速器

在短短几年，英伟达™ Fermi GPU 增强了世界上一些速度最快的超级计算机以及全球数以万计的研究集群。 目前，有了新 Kepler GK110 GPU，NVIDIA 又一次提高了 HPC 行业的标准。Kepler GK110 GPU 由 71 亿个晶体管组成，是创造的一个工程奇迹，解决了 HPC 行业中最严峻挑战。 Kepler 设计的初衷就是利用卓越的电源效率达到计算性能的最大化。该架构的创新之处在于使混合计算大大简化，适用于更广泛的应用，更容易获得。Kepler GK110 GPU 是计算主力，具有每秒万亿次整数，单精度，双精度浮点运算性能和最高的内存带宽。第一个以 GK110 为基础的产品将是 Tesla K20 GPU 计算加速器。

本技术简介目的在于快速汇总三个 Kepler GK110 中最重要的特点 GPU: SMX、Dynamic Parallelism 和 Hyper-Q。 有关其他架构特点的更多详细信息，请参考 Kepler GK110 白皮书。

图 2：SMX: 192 个 CUDA 核、32 个特殊功能单元 (SFU) 和 32 个加载/存储单元 (LD/ST)

SMX - 新一代流式多处理器

Kepler GK110 GPU 的核心是 SMX 单元，集成了几个架构创新，这不仅使其成为有史以来功能最强大的流式多处理器（SM），而且还最省电、最具编程性。

DyNAMIC PARALLELISM — 动态创建工作

在设计 Kepler GK110 架构的总体目标之一是使开发人员更容易更轻松地利用 GPU 的巨大并行处理能力。

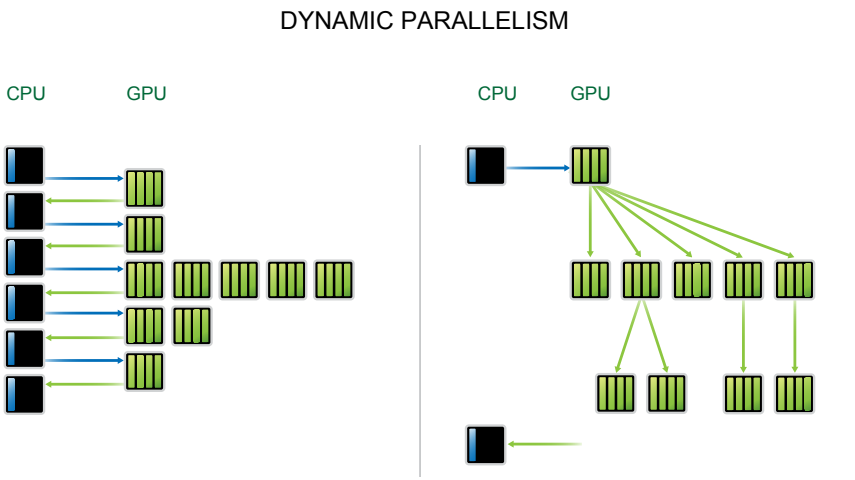

图 3：没有 Dynamic Parallelism 的情况下，CPU 启动 GPU 上的每个内核。有了该新功能的情况下，Kepler GK110 GPU 现在可以启动嵌套内核，不需要与 CPU 进行通信。

为此，新的 Dynamic Parallelism 功能，使 Kepler GK110 GPU 能通过应用不返回主机 CPU 的数据而动态创建新线程。 这能使多个程序有效地直接在 GPU 上运行，因为内核现在有能力独立承担所需的额外工作量。

任何内核可以启动另一个内核，并创建处理额外的工作所需的必要流程、事件和依赖，而无需主机 CPU 的介入。 这种简化的编程模式更易于创建、优化和维护。 它还通过为 GPU 维持与传统 CPU 内核启动工作负载相同的语法，创建了一个程序员友好环境。

Dynamic Parallelism 拓宽了目前在各领域利用 GPU 可以完成的应用程序。 应用程序可以动态启动中小型并行工作负载，这在以前是非常昂贵的。

Hyper-Q 在基于 MPI 并行计算机系统中使用会有明显的优势。 通常会在为多核 CPU 系统中运行而创建基于 MPI 的传统算法。 由于以 CPU 为基础的系统可以有效处理的工作负载通常比使用的 GPU 处理的较小，所以一般每个 MPI 进程中通过的工作量是不足以完全占据 GPU 处理器。

HyPER-Q —最大化 GPU 资源

Hyper-Q 允许多个 CPU 核同时在单一 GPU 上启动工作，从而大大提高了 GPU 的利用率并削减了 CPU 空闲时间。此功能增加了主机和 Kepler GK110 GPU 之间的连接总数，允许 32 个并发、硬件管理的连接，与 Fermi 相比，Fermi 只允许单个连接。 Hyper-Q 是一种灵活的解决方案，允许 CUDA流程和消息传递接口（MPI）进程的连接，甚至是进程内的线程的连接。先前被假依赖限制的现有应用程序，可以在不改变任何现有代码的情况下，达到 32 倍的性能提升。

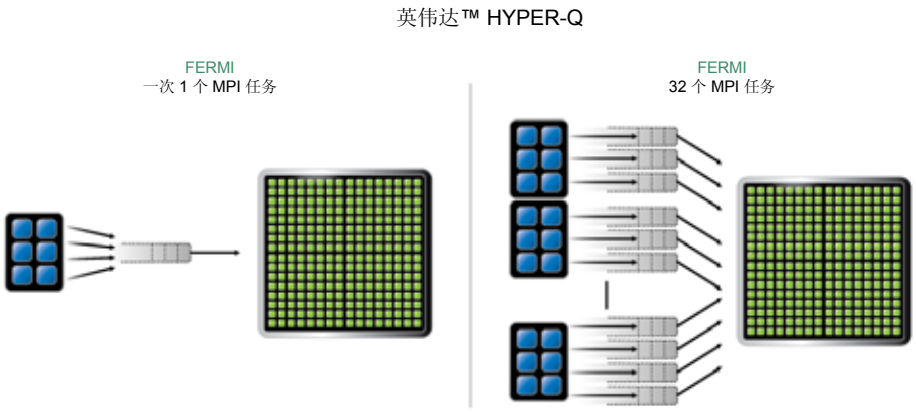

图 4：Hyper-Q 允许使用一个单独的工作队列同时运行所有流程。 在 Fermi 模式下，由于单一的硬件工作队列引起的流程内的依赖，并发受限。

虽然可以一直发出多个 MPI 进程同时运行在 GPU 上，但是这些进程有可能由于假依赖会成为瓶颈，迫使 GPU 低于最高效率地运行。 Hyper-Q 消除了假依赖的瓶颈，并大幅提高了从系统 CPU 将 MPI 进程移动到 GPU 的处理速度。

Hyper-Q 必定会是 MPI 应用程序性能提高的驱动。

结束语

Kepler GK110 GPU 进行了工程设计，提供具有卓越电源效率的开创性性能，而使 GPU 较之前更易于使用。 SMX、Dynamic Parallelism 和 Hyper-Q 是 Kepler GK110 GPU 中三项重要的创新，为我们的客户带来这些现实的好处。 有关其他架构特点的更多详细信息，请参考 Kepler GK110 白皮书，网址是 http://www.nvidia.com/object/nvidia-kepler.html。

如需了解有关英伟达™（NVIDIA®）Tesla的更多信息，敬请访问www.nvidia.com/tesla。