TESLA® CASE STUDY

CUDA 成为多业务领域的推进器

背景

计算技术的迅猛发展为其它科学应用领域提供了更好的平台。而早先借助于摩尔定律的发展变化，人们只要等待时间的推进便可加速其相关应用的计算效率。但任何事物都有一定限制，摩尔定律随着时间的推进也渐入瓶颈期。而随之而来的便是并行计算技术的产生。同时，早先被用来做视频图形的 GPU 反而也在并行计算相关领域发挥了其独特的优势。

核数众多是 GPU 的一大亮点，同时 Nvidia 公司率先利用了基于 C 平台的并行计算编程模型 CUDA，促进了代码并行移植至 GPU 的可靠性。因而 CUDA 并行编程模型自 2007 年开始就以极大的势头迅速的在计算机以及其它相关领域席卷开来。

挑战

在许多业务领域中，面对大规模计算，其最大的特点可以总结如下：

1. 重复计算量大，单一计算形式简单
2. 由于计算尺度量大，网格规模大，导致计算时间较长
3. 单节点 CPU 计算能力强，但并行规模未能符合要求；多 CPU，多节点并行投入资金量大，耗电量大，占地规模也相对较大。

方案

曙光 GPU 服务器以其独有的优势，成为 NVIDIA GPU 编程的重要载体。 通过曙光的 GPU 服务器的各类方案设计，CUDA 也充分发挥了其优势，从最早期的简单的 CUDA 并行，到后来 GPU 之间的 NvLink 连接，使得 GPU 之间进行 P2P 传递数据而不经过 CPU。CUDA 也为各类计算任务带了极大的便捷编程模式。

单以计算流体力学中的一些基本数值格式为例，许多格式的计算需要计算网格中每个单元对应的流体密度、压强、动量等，这些物理量的计算需要涉及目标单元的邻居单元，而在区域分块的时候还需要相互传递边界虚单元的值。CUDA 可以将这些物理量的值传递到 GPU 上，在 GPU 上进行每个物理量的更新计算，并借助 P2P 函数传递边界虚单元。

同时由于 GPU 上面有各类内存，如全局显存，共享显存，寄存器变量缓存，纹理显存等。而全局内存和共享内存的访问时钟周期还大有差异。这样我们可以充分利用共享显存的访问速度快的特点对各类问题进行并行优化设计。而在 CUDA 代码的设计方面，不同的优化方式可以得到不同的运行速度，这也使得很多学者在不断的修改优化着自己的代码，很多技术人员也乐于去 PK 他们的代码。

影响

CUDA 的代码是基于 C 的，而 C 语言是目前为止全世界最流行最通用的语言，CUDA 就是在 C 的基础上编写了一套模型函数，而函数的名称都与 C 有相对应的形式。

为了迎合科学计算等领域，CUDA 还衍生了 CUDA Fortran。加大了 CUDA 和 GPU 的应用开发平台。正是这样，许多工程上的应用也借助 CUDA 将其编写到其模块中，像 Ansys 等 CFD 应用软件都已经添加了 GPU 的计算部分。足可见基于 GPU 的 CUDA 编程模型在软件应用方面的促进作用。