

课程实验报告

课 程 名 称： 计算机系统设计

实验项目名称： Tesla GPU和GeForce GPU的比较

专 业 班 级：物联1702

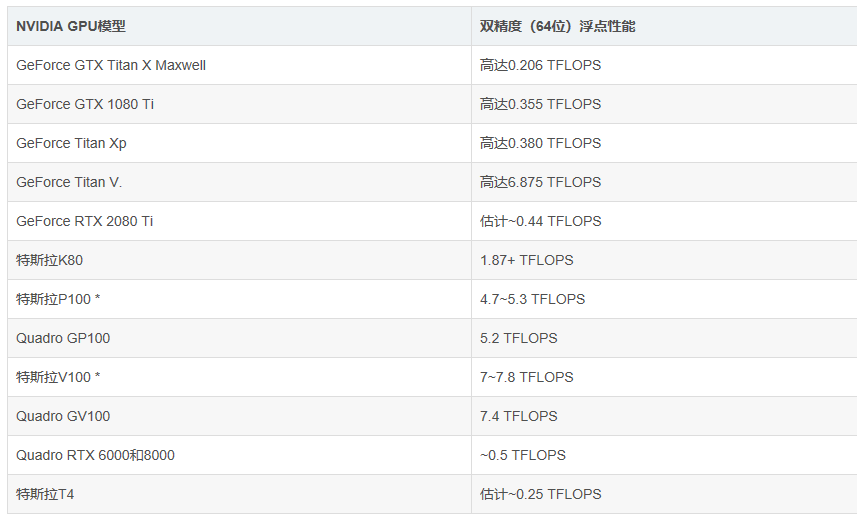
姓 名：易纪千

学 号： 201708010618

完 成 时 间： 2019年12月 22 日

信息科学与工程学院

### **一、FP64 64位（双精度）浮点计算**

GeForce和Tesla / Quadro GPU之间双精度浮点计算性能的比较：

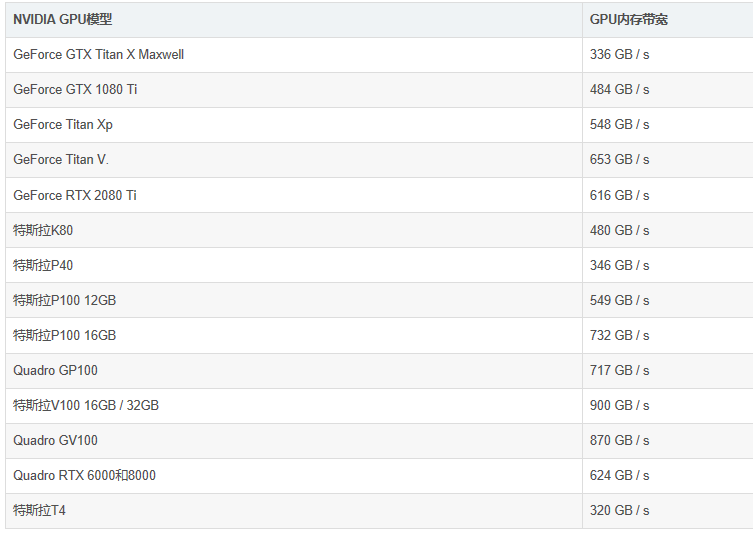
1. **FP16 16位（半精度）浮点计算**

尽管所有NVIDIA“Pascal”和后来的GPU都支持FP16，但在许多以游戏为中心的GPU上，性能显着降低。以下是GeForce和Tesla / Quadro GPU之间半精度浮点计算性能的比较：



1. **GPU内存性能**

计算密集型应用程序需要高性能计算单元，但快速访问数据也很关键。对于许多HPC应用程序，除非内存性能也得到改善，否则计算性能的提高无济于事。因此，Tesla GPU提供比GeForce GPU更好的实际性能：



1. **GPU内存大小**

通常，系统运行的内存越多，运行的速度就越快。对于某些HPC应用程序，除非有足够的内存，否则甚至无法执行单次运行。对于其他人来说，除非有足够的内存，否则结果的质量和保真度会降低。Tesla GPU提供的内存是GeForce GPU的两倍：



**五、DMA引擎**

GPU的直接内存访问（DMA）引擎允许在系统内存和GPU内存之间进行快速数据传输。由于此类传输是任何实际应用程序的一部分，因此性能对GPU加速至关重要。缓慢的传输导致GPU核心处于空闲状态，直到数据到达GPU内存。同样，慢速返回会导致CPU等待，直到GPU完成返回结果。

GeForce具有单个DMA引擎，可以一次传输一个方向的数据。如果正在将数据上载到GPU，则在上载完成之前，无法返回由GPU计算的任何结果。同样，从GPU返回的结果将阻止需要上传到GPU的任何新数据。Tesla GPU采用双DMA引擎来缓解这一瓶颈。数据可以同时传输到GPU和GPU中。