

课程实验报告

课 程 名 称： 计算机系统设计

专 业 班 级： 物联1702

姓 名： 龚竞舟

学 号： 201708010612

完 成 时 间： 2019 年 12 月 5 日

计算机工程系

## 实验题目

Tesla GPU架构分析

1. **实验内容**
2. **NVIDIA Tesla V100 GPU架构简介**

自10年前推出开创性的CUDA GPU计算平台以来，NVIDIA GPU的每一次更新换代都将应用程序性能和能效推升到更高的水平，同时还新增若干重要的计算功能，并简化GPU编程难度。如今NVIDIA GPU为数千个高性能计算（HPC）、数据中心和机器学习应用程序提供加速动力，已成为推动人工智能革命的领先计算引擎。

全新的NVIDIA Tesla V100加速器搭载强大的新款Volta GV100 GPU。GV100不仅汲取了上一代产品Pascal GP100 GPU的精进之处，而且还显著提高性能和可扩展性，并新增许多能够改进可编程性的功能，这些改进将有效提升HPC、数据中心、超级计算机以及深度学习系统和应用程序的性能。本文将对Tesla V100 GPU架构进行介绍。

1. **NVIDIA Tesla V100主要特性**

NVIDIA Tesla V100加速器性能在全球并行处理器中看成出类拔萃，GV100 GPU包含211亿根晶体管，芯片大小为815mm^2，采用专为NVIDIA定制的全新TSMC 12 nm FFN高性能制造工艺精心打造而成。以下为Tesla V100部分主要计算特性：

1. 专为深度学习优化的全新流多处理器架构：

Volta GPU中央配备有全新设计的SM处理器架构，全新Volta SM的节能效率相较上一代Pascal产品提升50%，在同一功率电路下可显著提高FP32和FP64的性能。专为深度学习设计的新Tensor核心在训练方面可提供高达12倍的TFLOPS峰值。

1. 第二代NVIDIA NVLink高速互联功能提供更高宽带与更多链路。
2. 拥有经过重点调整的16GB HBM2内存子系统，可提供900GB/s的内存宽带峰值。
3. 拥有Volta多进程服务。

可为CUDA MPS服务器关键组件实现硬件加速，从而为共享GPU的多个计算应用程序提高性能，实现隔离并改进服务质量。

1. 统一内存寻址和地址转换服务质量提升，新的地址转换服务支持允许GPU直接读取CPU的分页表。
2. 最大性能模式和最大效率模式：最大性能模式下，加速器将以300W的TDP级别运行，为需要最快计算速度和最高数据吞吐量的应用程序加速。在最大效率模式下，数据中心管理员可调节Tesla V100的功率利用率。
3. 协作组和新的协作启动API：

协作组是CUDA 9中引入的新式编程模型，可用于组织线程通信群组。

1. 针对Volta优化的软件。

Volta优化版的GPU加速库如cuDNN / cuBLAS / TensorRT能充分利用Volta GV100架构的新功能。

1. **深入理解NVIDIA Tesla GV100 GPU硬件架构**

NVIDIA Tesla V100加速器配备Volta GV100 GPU，是世界领先的高性能并行计算处理器，与上一代Pascal GP100 GPU一样，GV100 GPU由多个GPU处理集群GPC、纹理处理集群TPC，流多处理器SM和内存控制器组成。完整的GV100 GPU包含以下组件：

1、6个GPC，每个GPC拥有7个TPC（各包含两个SM）

2、84个Volta SM，每个SM拥有64个FP32核心，64个INT32核心，32个FP64核心，8个Tensor核心，4个纹理单元。

3、8个512位内存控制器，共4096位。含有84个SM的完整GV100 GPU总共拥有5376个FP32核心，5376个INT32核心，2688个FP64核心，672个Tensor核心以及336个纹理单元。每个HBM2 DRAM堆栈由一对内存控制器控制。完整的GV100 GPU总共包含6144KB的L2缓存。

**下图是含84个SM的完整Tesla GV100 GPU**



1. **卓越的性能和能效**

NVIDIA GPU每经一次开发换代，性能和能效都会得到显著提升。在最大性能模式下，Tesla V100加速器将以300W的热设计功耗级别运行，为需要最快计算速度和最高数据吞吐量的应用程序加速。在最大效率模式下，数据中心管理员可调节Tesla V100的功率利用率。

最大效率模式允许数据中心管理员以最佳性能功耗比运行加速器，V100可设置以一定的功率/性能比运行，以在最佳性能和最高能效间取舍。