**Tesla GPU架构分析**

**智能1702班 尹淑青 201708010601**

1. Tesla GPU的简单介绍

NVIDIA的GPU产品主要有GeForce、Tesla和Quadro三大系列，虽然从硬件角度来看它们都采用同样的架构设计，也都支持用作通用计算(GPGPU)，但因为它们分别面向的目标市场以及产品定位的不同，这三个系列的GPU在软硬件的设计和支持上都存在许多差异。

其中Quadro的定位是专业用途显卡而Tesla的定位是专业的GPGPU，单价相对较高，也都很少会被用作其他用途。但面向消费者的GeForce显卡却因为出货量大，价格较低的缘故经常被当作另外两个专业产品的替代品来使用。 本文主要探讨Tesla。

1. NVIDIA之Tesla、GeForce和Quadro系列GPU对比

芯片及启用核心数量的差异：

虽然同一代的GPU产品都采用相同的核心架构，但不同型号的GPU采用的是不同等级的GPU核心，比如作为Tesla系列旗舰的P100采用的是GP100核心，而P40和P4则分别采用的是GP102和GP104核心，在GeForce系列产品里，则还有采用更低的GP106/107/108等型号核心的产品。



双精度浮点(FP64)计算性能的差异：

在采用Pascal架构的GPU核心里，只有GP100采用了单精度计算单元和双精度计算单元为2:1的比例设计，其他核心采用的都是32:1的设计，这也就造成采用GP100核心的GPU比如Tesla P100和在双精度计算能力方面远远超过其他型号。

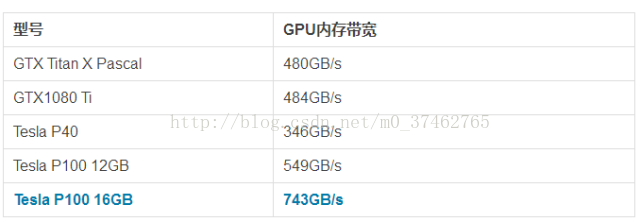
ECC内存的错误检测和纠正：

在运行3D游戏的GeForce显卡上，即使出现一些内存错误通常也不会造成什么严重的问题，对于个人用户来说，显示的画面偶尔出现些许的错误完全可以容忍甚至会被忽视。但对于计算领域来说，就非常依赖于GPU返回数据的准确性，即使内存出现单比特错误也可能导致最终计算结果的极大误差。

GeForce系列显卡不具备错误检测和纠正的功能，但Tesla系列GPU因为GPU核心内部的寄存器、L1/L2缓存和显存都支持ECC校验功能，所以Tesla不仅能检测并纠正单比特错误也可以发现并警告双比特错误，这对保证计算结果的准确性来说非常重要。

GPU内存性能：

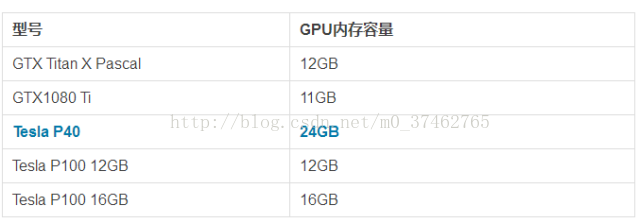
计算密集型应用程序不仅需要GPU提供高性能计算单元，也需要GPU提供快速访问数据的能力，否则再好的GPU核心也将成为巧妇难为无米之炊。 对于许多HPC应用程序，GPU内存性能的差异对最终结果的影响甚至比计算能力更明显，Tesla GPU可以提供比GeForce GPU更好的内存带宽：



造成这种性能差异的主要原因是GeForce GPU使用GDDR5或GDDR5x内存，而Tesla P100 GPU则使用性能更好的HBM2内存。

GPU内存(显存)容量：

一般来说，系统的内存越多，运行速度越快。尤其是对于某些HPC应用程序来说，内存不够时甚至不能执行单次运行。GeForce 显卡最大只能提供12GB的显存，而Tesla P40 GPU则最大可以提供2倍的显存——24GB，这对GPU执行深度学习运算时使用更大的框架提供了支持。



Pascal核心的Tesla的统一内存技术还允许GPU共享彼此的内存以加载更大的数据集。

1. Tesla V100介绍

2017年5月11日，正式发布了VIDIA Tesla V100的Volta架构GPU，在当时，Tesla V100是史上规模最庞大的GPU，拥有超过210亿个晶体管，是上代Tesla P100的1.37倍，核心面积达到了创纪录的815平方毫米。此外，Tesla V100还增加与深度学习高度相关的Tensor单元，Tensor性能号称可以达到120TFLOPS。同时，Tesla V100的二级缓存及寄存器大小也有所增加，L2缓存由Tesla P100的4MB增加到了6MB，每组SM单元的寄存器文件大小总数从14MB增加到了20MB。NVIDIA表示，Tesla V100将首先用在用于深度学习超算DGX-1上，内部拥有8张Tesla V100计算卡，峰值计算性能高达960TFLOPS，号称用8个小时就能完成TITANX八天的工作量。

目前NVIDIA官网发布的Tesla V100消息

Tesla V100拥有640个[Tensor内核](https://www.nvidia.cn/data-center/tensorcore/)，是世界上第一个突破100万亿次(TFLOPS) 深度学习性能障碍的GPU。新一代NVIDIA NVLink以高达300 GB/s的速度连接多个V100GPU，在全球打造出功能极其强大的计算服务器。

Tesla V100的设计能够融合人工智能和高性能计算。它为高性能计算系统提供了一个平台，在用于科学模拟的计算机科学和用于在数据中发现见解的数据科学方面表现优异。通过在一个统一架构内搭配使用NVIDIA CUDA内核和Tensor内核，配备Tesla V100 GPU的单台服务器可以取代数百台仅配备通用CPU的服务器来处理传统的高性能计算和人工智能工作负载。

1. NVIDIA CUDA内核和Tensor内核

CUDA内核：

CUDA（Compute Unified Device Architecture），是显卡厂商NVIDIA推出的运算平台。 CUDA是一种由NVIDIA推出的通用并行计算架构，该架构使GPU能够解决复杂的计算问题。 它包含了CUDA指令集架构（ISA）以及GPU内部的并行计算引擎。 开发人员现在可以使用C语言来为CUDA架构编写程序，C语言是应用最广泛的一种高级编程语言。所编写出的程序可以在支持CUDA的处理器上以超高性能运行。CUDA3.0已经开始支持C++和FORTRAN。

[Tensor内核](https://www.nvidia.cn/data-center/tensorcore/)（剖析Tensor Core）

在关于Volta混合精度Tensor Core的几个谜团中，一个比较烦人的问题是4×4矩阵乘法的能力。Tensor Core是一种新型处理核心，它执行一种专门的矩阵数学运算，适用于深度学习和某些类型的HPC。Tensor Core执行融合乘法加法，其中两个4\*4 FP16矩阵相乘，然后将结果添加到4\*4 FP16或FP32矩阵中，最终输出新的4\*4 FP16或FP32矩阵。

NVIDIA将Tensor Core进行的这种运算称为混合精度数学，因为输入矩阵的精度为半精度，但乘积可以达到完全精度。碰巧的是，Tensor Core所做的这种运算在深度学习训练和推理中很常见。

Tensor Core虽然在GPU里是全新的运算单元，但其实它与标准的ALU流水线并没有太大差别，只不过Tensor Core处理的是大型矩阵运算，而不是简单地单指令流多数据流标量运算。Tensor Core是灵活性和吞吐量权衡的选择，它在执行标量运算时的表现很糟糕，但它可以将更多的操作打包到同一个芯片区域。

Tensor Core虽然有一定的可编程性，但仍然停留在4\*4矩阵乘法累加层面上，并且不清楚累积步骤是如何以及何时发生的。尽管被描述为进行4\*4矩阵数学运算，但实际上Tensor Core运算似乎总是使用16\*16矩阵，并且操作一次跨两个Tensor Core进行处理。这似乎与Volta架构中的其他变化有关，更具体地说，与这些Tensor Core是如何集成进SM中有关。

对于Volta架构，SM被划分为四个处理块或子核。对于每个子核，调度器每个时钟向本地分支单元（BRU）、Tensor Core阵列、数学分派单元或共享MIO单元发出一个warp指令，这就首先阻止了Tensor运算和其他数学运算同时进行。在利用两个Tensor Core时，warp调度器直接发出矩阵乘法运算，并且在从寄存器接收输入矩阵之后，执行4\*4\*4矩阵乘法。待完成矩阵乘法后，Tensor Core再将得到的矩阵写回寄存器。

在Tensor Core执行实际指令时，即使在使用NVVM IR（LLVM）的编译器级别上，也仅存在用于warp级矩阵操作的本征，对于CUDA++和PTX ISA，warp级别仍然是唯一级别。加载输入矩阵的形式是每个扭曲线程持有一个片段，其分布和身份均未指定。从广义上讲，它遵循标准CUDA核心的基于线程级别拼接的GEMM计算的相同模式。

从NVIDIA的角度来看，Volta不是一颗深度学习的专用ASIC，它仍然覆盖GPGPU的领域，因此保持CUDA可编程Tensor Core适用于GEMM / cuBLAS和HPC是合乎逻辑的。对于CUDA c++的CUTLASS来说，情况更是如此，因为它的WMMA API支持旨在为广泛的应用程序启用Tensor CoreGEMM操作。从根本上说，NVIDIA深度学习硬件加速的发展与cuDNN（以及cuBLAS）的发展有很大关系。