湖南大學

HUNAN UNIVERSITY

《 计算机系统设计 》

题 目 Tesla GPU 架构分析

专 业 智能科学与技术

班 级 智能 1702

姓 名 唐昌均

学 号 201708010712

1.简介

Tesla GPU 的 20 系列产品家族基于代号为“Fermi”的下一代 CUDA 架构，支持技术与企业计算所“必备”的诸多特性，其中包括 C++支持、可实现极高精度与可扩展性的 ECC 存储器以及 7 倍于 Tesla 10 系列 GPU 的双精度性能。Tesla? C2050 与 C2070 GPU 旨在重新定义高性能计算并实现超级计算的平民化。

与最新的四核 CPU 相比，Tesla C2050 与 C2070 计算处理器以十分之一的成本和二十分之一的功耗即可实现同等超级计算性能。

2.特性

基于新一代 Fermi CUDA 架构的 GPU：与基于最新四核 CPU 的纯 CPU 系统相比，该 GPU 以十分之一的成本和二十分之一的功耗即可实现同等的集群性能。

448 个 CUDA 核心：每颗 GPU 最高可实现 515 Gigaflop 双精度峰值性能，从而让一台工作站即可实现 Teraflop 级甚至更高的性能。每颗 GPU 的单精度峰值性能超过 1 Teraflop。

ECC 存储器：能够满足工作站计算精度与可靠性方面的关键需求。能够为存储器中的数据提供保护功能，从而为应用程序增强数据完整性和可靠性。寄存器文件、L1/L2 高速缓存、共享存储器以及 DRAM 均受 ECC 的保护。

台式机上的集群性能：与一个小型服务器集群相比，配备多颗 GPU 的单台工作站能够更快地解决大型难题。

每颗 GPU 最多配备 6GB GDDR5 存储器：更大的数据集能够保存在直接附属于 GPU 的本地存储器上，从而实现了性能的最大化并减少了数据传输的情况。

NVIDIA®（英伟达？）并行 DataCache?：能够为物理效果解算器、光线追踪以及稀疏矩阵乘法等诸多算法加速，在这些算法中，数据地址事先都是未知的。每个流式多处理器模块均包含一个可配置的 L1 高速缓存，所有处理器核心使用统一的 L2 高速缓存。

NVIDIA®（英伟达？）GigaThread?引擎：通过多项技术实现了吞吐量的最大化，其中包括 10 倍于上一代架构的高速上下文切换、并发内核执行以及改良的线程块调度。

异步传输：计算核心在 PCIe 总线上传输数据的同时还能够处理其它数据，因而增强了系统性能。即便是地震处理这类需要大量数据传输的应用程序，也能够通过事先将数据传输至本地存储器的方法来最大限度提升计算效率。

CUDA 编程环境受到各种编程语言与 API 的广泛支持：开发人员无论选择 C 语言、C++、OpenCL、DirectCompute 还是选择 Fortran 语言，都能够实现应用程序的并行机制，进而利用“Fermi”GPU 的创新架构。Microsoft Visual Studio 开发人员可以使用 NVIDIA®（英伟达？）Parallel Nsight 工具。

高速 PCIe Gen 2.0 数据传输率：实现了主系统与 Tesla 处理器之间带宽的最大化。让 Tesla 系统能够应用于几乎所有具备一条开放式 PCIe x16 插槽且符合 PCIe 规范的主系统。

3.实例

Tesla GPU 有很多系列，很多版本，所以架构也有很多种。这里我们就其中一种来分析。

Tesla V100：AI 计算和 HPC 的动力之源

Tesla V100 加速器性能在全球并行处理器中堪称出类拔萃，设计目的在于处理计算量巨大的 HPC、人工智能和图形工作负载。

V100 包含了 211 亿根晶体管，芯片大小为 815 平方毫米。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。相比上一代 Pascal GPU GV100 的计算性能显著提高 GPU 资源利用率。GV100 还可 GPU 资源利用率。GV100 是一款极为节能的处理器，可实现出色的性能功耗比。

主要特性：

A．专为深度学习优化的全新流多处理器（SM）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。专为深度学习设计新 Tensor 核心在训练方面可提供高达 12 倍的 TFLOPS 峰值，而在推理方面则可提供 6 倍的 TFLOPS 峰值。此外，通过使用单独的并行整数和浮点数据路径，Volta SM 在处理器包含计算和寻址计算的混合工作负载时也更为高效。Volta SM 新式独立线程调度功能，可在并行线程之间实现更精细的同步与合作。最后是 L1 数据缓存和共享内存单元的全新组合，在大幅提升性能之余更简化了编程。

B．HBM2 内存：高速、高效

Volta 拥有经重点调整的 16 GB HBM2 内存子系统，可提供 900GB/s 的内存带宽峰值。新一代 Samsung HBM2 内存与新一代 Volta 内存控制器的结合，能够提供比 Pascal GP100 高 1.5 倍的内存带宽，而运行多工作负载时的内存带宽利用率可达 95%。

C．Volta 多进程服务

Volta 多进程服务是 Volta GV100 架构的新功能，可为 CUDA MPS 服务器的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。此外，Volta MPS 还使 MPS 客户端的最大数量增至 3 倍，从 Pascal 时的 16 个增加到 Volta 的 48 个。

D．统一内存寻址和地址转换服务质量提升

GV100 统一内存寻址技术包含新的存取计数器，可更准确地将内存分页至对其读取最为频繁的处理器，同时提升处理器间共享内存范围的效率。

E．最大性能模式和最大效率模式

最大性能模式下，Tesla V100 加速器将以 300w 的 TDP（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 加速器的功率利用率，是加速器以最佳性能功耗比运行。可为同机架的所有 GPU 设置功率上限，从而大幅降低功耗，并使机架设备保持出色的性能。

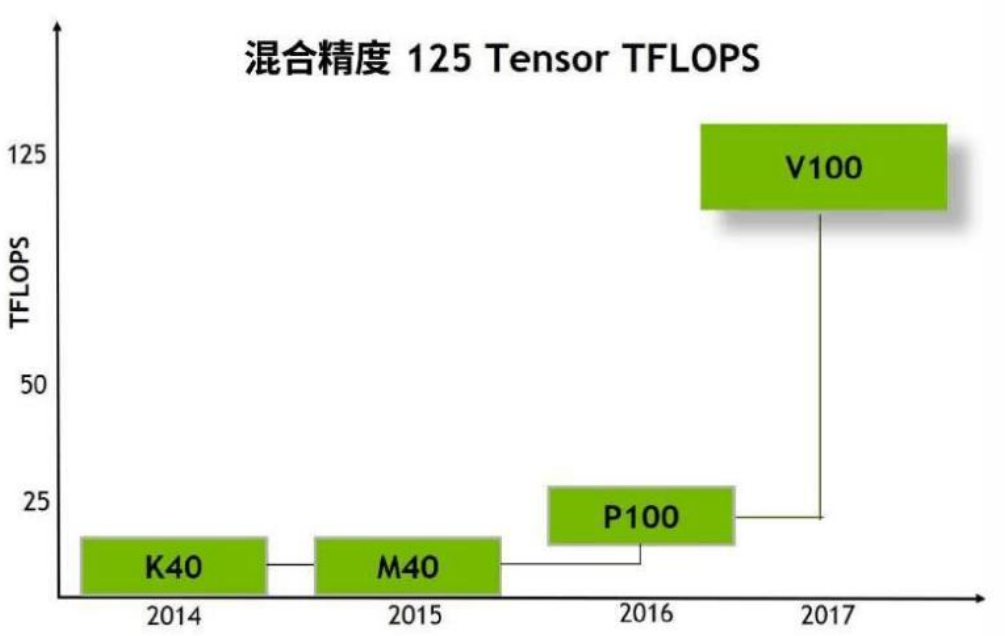
F．协作组和新的协作启动 API

协作组是 CUDA9 中引入的新式编程模型，可用于组织线程通信群组。协作组允许开发者表示线程通信粒度，帮助他们表达更丰富、更高效的并行分解方法。

G．针对 Volta 优化的软件

Caffe2/MXNet/CNTK/TensorFlow 等深度学习框架新版本以及其他框架皆可发挥出 Volta 的强大功能，缩短训练时间并获得更高的多节点训练性能。Volta 优化版的 GPU 加速库能够充分利用 Volta V100 架构的新功能，为深度学习推理和高性能计算应用程序提供高性能。

Tesla V100 可提供行业领先的浮点和整数性能。下图是计算速率峰值。



Tesla 各系列 GPU 性能比较。

Tesla 产品	Tesla K40	Tesla M40	Tesla P100	Tesla V100
GPU	GK180 (Kepler)	GM200 (Maxwell)	GP100 (Pascal)	GV100 (Volta)
SM 数量	15	24	56	80
TPC 数量	15	24	28	40
FP32 核心数/SM	192	128	64	64
FP32 核心数/GPU	2880	3072	3584	5120
FP64 核心数/SM	64	4	32	32
FP64 核心数/GPU	960	96	1792	2560
Tensor 核心数/SM	NA	NA	NA	8
Tensor 核心数/GPU	NA	NA	NA	640
GPU 加速频率	810/875 MHz	1114 MHz	1480 MHz	1530 MHz
FP32 TFLOPS峰值 [1]	5	6.8	10.6	15.7
FP64 TFLOPS峰值 [1]	1.7	0.21	5.3	7.8
Tensor TFLOPS峰值 [1]	NA	NA	NA	125
纹理单元数量	240	192	224	320
显存位宽	384 位 GDDR5	384 位 GDDR5	4096 位 HBM2	4096 位 HBM2
显存容量	最大为 12 GB	最大为 24 GB	16 GB	16 GB
L2 缓存大小	1536 KB	3072 KB	4096 KB	6144 KB
共享内存大小/SM	16 KB/32 KB/48 KB	96 KB	64 KB	最大可配置为 96 KB
寄存器文件大小/SM	256 KB	256 KB	256 KB	256KB
寄存器文件大小/GPU	3840 KB	6144 KB	14336 KB	20480 KB
TDP（热设计功耗）	235 W	250 W	300 W	300 W
晶体管数量	71 亿	80 亿	153 亿	211 亿
GPU 芯片大小	551 mm²	601 [2]	610 mm²	815 mm²
制造工艺	28 nm	28 nm	16 nm FinFET+	12 nm FFN