南開大學
Nankai University

计算机学院

并行程序设计作业 1

并行计算机体系结构调研

姓名：林语盈

学号：2012174

专业：计算机科学与技术

2022 年 3 月 2 日

目录

1 并行计算机发展历史简述

最近几十年间，计算机的计算能力遵从摩尔定律迅速增长。[1] 2002 年以前，计算机时钟频率的增加使计算机算力增加，但这带来了功率的问题，计算机的功率随之急速上升，计算机过热的问题无法解决。于是，2002 年以后，计算机的计算能力提高主要依赖于多核、并行的架构。

GPU，即图形处理器。[2]，原仅用于图形渲染和显示任务，相比于 CPU，其具有集成大量计算单元的特点。随着 AI 的发展，深度学习等新的计算任务的出现，GPU 的大量计算单元被用于计算矩阵运算等重复性较强的任务，并行程序框架也随之发展，GPU 的可编程性不断增强。现在，GPU 的应用能力已经远远超出了图形处理器，利用 GPU 完成通用计算的研究逐渐活跃起来，将 GPU 用于图形渲染以外领域的计算成为 GPGPU(General Purpose computing on graphics processing units，基于 GPU 的通用计算)

NVIDIA 公司在 1999 年发布 Geforce256 图形处理芯片时首先提出 GPU 的概念。从此 NVIDIA 显卡的芯就用 GPU 来称呼。GPU 使显卡减少了对 CPU 的依赖，并进行部分原本 CPU 的工作，尤其是在 3D 图形处理时。20 年间，NVIDIA 生产了包括 Kepler、Maxwell、pascal、Turing、Volta 架构的 GPU 芯片。

NVIDIA计算卡进化历程

	K40	M40	P100	V100	A100
发布时间	2013.11	2015.11	2016.4	2017.05	2020.05
架构	Kepler	Maxwell	Pascal	Volta	Ampere
制程	28 nm	28 nm	16 nm	12 nm	7nm
晶体管数量	71亿	80亿	153亿	211亿	510亿
Die Size	551 mm²	601 mm²	610 mm²	815 mm²	826 mm²
最大功耗	235 W	250 W	300 W	300 W	400 W
Streaming Multiprocessors	15	24	56	80	108
Tensor Cores	NA	NA	NA	640	432
FP64 CUDA Cores	960	96	1792	2560	3456
FP32 CUDA Cores	2880	3072	3584	5120	6912
FP32 峰值算力	5.04 TFLOPS	6.08 TFLOPS	10.6 TFLOPS	15.7 TFLOPS	19.5 TFLOPS
稀疏Tensor Core F32 峰值算力	NA	NA	NA	NA	312 TFLOPS

图 1.1: 英伟达 GPU 架构发展历程

2020 年 5 月 14 日，NVIDIA 创始人兼首席执行官黄仁勋在 NVIDIA GTC 2020 主题演讲中介绍了基于最新 Ampere 架构的 NVIDIA A100 GPU。[3]

2 英伟达安培

2.1 英伟达安培总体结构

NVIDIA GA100 GPU 由多个 GPU 处理群集 (GPC)，纹理处理群集 (TPC)，流式多处理器 (SM) 和 HBM2 显存控制器组成。

GA100 GPU 的完整实现包括以下单元：

- 每个完整 GPU 有 8 个 GPC，每个 GPC 有 8 个 TPC，每个 TPC 有 2 个 SM，每个 GPC 有 16 个 SM，总共 128 个 SM；

- 每个 SM 有 64 个 FP32 CUDA 核，总共 8192 个 FP32 CUDA 核；
- 每个 SM 有 4 第三代 Tensor Core，总共 512 个第三代 Tensor Core；
- 总共 6 个 HBM2 堆栈，12 个 512 位显存控制器；

基于 GA100 GPU 实现的 A100 Tensor Core GPU 包括以下单元：

- 每个 GPU 有 7 个 GPC，每个 GPC 有 7 个或 8 个 TPC ，每个 TPC 有 2 个 SM，每个 GPC 最多 16 个 SM，总共 108 个 SM；
- 每个 SM 有 64 个 FP32 CUDA 核，总共 6912 个 FP32 CUDA 核；
- 每个 SM 有 4 个第三代 Tensor Core，总共 432 个第三代 Tensor Core；
- 总共 5 个 HBM2 堆栈，10 个 512 位显存控制器；

128 个 SM 的完整 GA100 GPU 架构，如图2.2所示。

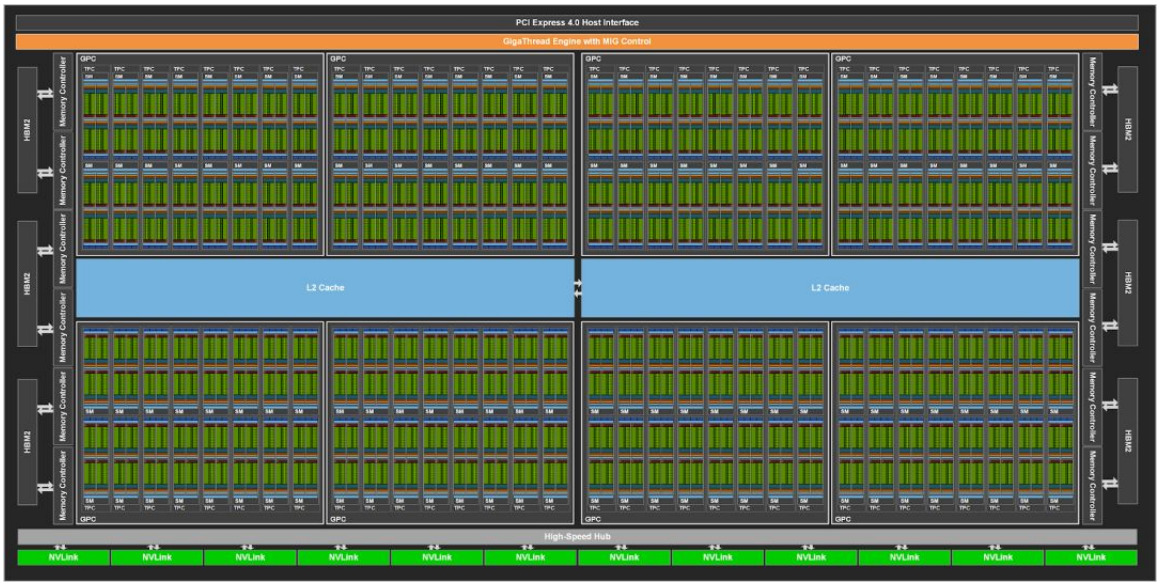

图 2.2: 128 个 SM 的完整 GA100 GPU

A100 SM 引入新的第三代 Tensor Core，每个 Tensor Core 在每个时钟周期内可执行 256 个 FP16/FP32 FMA 计算。

A100 每个 SM 有 4 个 Tensor Core，每个 SM 提供 1024 个密集的 FP16/FP32 FMA 操作，与 Volta 和 Turing 相比，每个 SM 的算力翻倍。

GA100 GPU 的 SM 架构，如图2.5所示。

图 2.3: 128 个 SM 的完整 GA100 GPU

2.2 缓存结构

A100 L2 cache 可提供的读取带宽是 V100 的 2.3 倍，因而能以比从 HBM2 内存读写高得多的速度缓存和重复访问更大的数据集和模型。L2 cache residency control 被用于优化容量利用率，可以管理数据以保存或从缓存中删除数据。

为了提高效率和增强可扩展性，A100 增加了计算数据压缩，可节省高达 4 倍的 DRAM 读/写带宽、4 倍的 L2 读带宽和 2 倍的 L2 容量。

此外，NVIDIA 通过将 L1 cache 和 shared memory 单元结合到一个内存块的方式来提高内存访问的性能，同时简化了编程和调优步骤，并降低软件的复杂性。

每个 SM 中的 L1 cache 和 shared memory 单元总容量达 192 KB，是此前 V100 的 1.5 倍。

2.3 Tensor Core 结构

第三代 Tensor Core 特性：

- 支持的数据类型有 FP16、BF16、TF32、FP64、INT8、INT4 和 INT1；
- 利用深度学习网络的细粒度结构化稀疏性，其性能相比标准 Tensor Core 计算翻倍；

- A100 中的 TF32 Tensor Core 计算可在 DL 和 HPC 中容易地加速 FP32 输入/输出数据，比 V100 FP32 FMA 快 10 倍，稀疏优化的情况下达到 20 倍；
- FP16/FP32 Tensor Core 混合精度计算为 DL 提供了更强的性能，是 V100 的 2.5 倍，稀疏优化下达到 5 倍；
- BF16/FP32 混合精度和 FP16/FP32 在同一速率下运行；
- FP64 Tensor Core 为 HPC 提供前所未有的双精度处理性能，速度是 V100 FP64 DFMA 的 2.5 倍；
- 带稀疏优化的 INT8 Tensor Core DL 推理速度是 V100 INT8 的 20 倍；
- 每个 SM shared memory / L1 缓存共有 192 KB，是 V100 SM 的 1.5 倍;
- 新的异步拷贝指令能够从 global memory 中将数据直接加载到 SM shared memory，可选地绕过 L1 缓存，降低中间寄存器堆（RF）的需求；
- 与异步拷贝指令搭配的异步 barrier；
- 新的 L2 缓存管理和驻留控制指令；
- 新的 warp 级别规约指令（由 CUDA 协作组提供支持）；
- 许多可以减少软件复杂度的编程性改进；

  A100 提供了 IEEE 兼容的 FP64 Tensor Core 计算，相比 V100 FP64 有 2.5 倍性能提升。A100 上新的双精度矩阵乘法指令取代了 V100 上的 8 条 DFMA 指令，减少了取指令、调度开销、寄存器读取、数据传输和 shared memory 读取带宽。

  A100 每个 SM 的计算次数为 64 FP64 FMA/clock（或 128 FP64 FMA/clock），是 V100 的 2 倍。拥有 108 个 SM 的 A100 可以最多再 FP64 下达到 19.5 TFLOPS，是 V100 的 2.5 倍。有了这些新格式的支持，A100 可以被用于 HPC 任务，迭代求解器以及其他的 AI 算法。

  下表显示了 A100 比 V100 的速度提升（TC=Tensor Core，GPU 在各自的时钟速度下）。

	V100	A100	A100 Sparsity	A100 Speedup	A100 Speedup with Sparsity
FP16	31.4 TFLOPS	78 TFLOPS	N/A	2.5x	N/A
FP16 with Tensor Core	125 TFLOPS	312 TFLOPS	624 TFLOPS	2.5x	5x
A100 BF16 TC vs.V100 FP16 TC	125 TFLOPS	312 TFLOPS	624 TFLOPS	2.5x	5x
FP32	15.7 TFLOPS	19.5 TFLOPS	N/A	1.25x	N/A
A100 TF32 TC vs. V100 FP32	15.7 TFLOPS	156 TFLOPS	312 TFLOPS	10x	20x
FP64	7.8 TFLOPS	9.7 TFLOPS	N/A	1.25x	N/A
A100 FP64 TC vs. V100 FP64	7.8 TFLOPS	19.5 TFLOPS	N/A	2.5x	N/A
A100 INT8 TC vs. V100 INT8	62 TOPS	624 TOPS	1248 TOPS	10x	20x
A100 INT4 TC	N/A	1248 TOPS	2496 TOPS	N/A	N/A
A100 Binary TC	N/A	4992 TOPS	N/A	N/A	N/A

图 2.4: A100 与 V100 的性能对比

2.4 CUDA

CUDA(Compute Unified Device Architecture)[4],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。开发人员可以使用 C 语言来为 CUDA架构编写程序，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。

与之相配，NVIDIA 近日了发布全新的 CUDA11 版本工具包，特别为新诞生的安培架构进行了优化。CUDA 11 完全支持在安培新架构上进行开发，包括 A100 GPU，以及基于它的 DGX A100、HGX A100 等多路系统，并支持安培架构的第三代 Tensor 张量核心，可针对不同数据类型加速混合精度矩阵计算，比如 TF32、Bfloat16。

而英伟达安培新的多实例 GPU （MIG ）特性允许 A100 张量核心 GPU 安全地划分为多达七个独立的 GPU 实例，用于 CUDA 应用程序，为多个用户提供单独的 GPU 资源，以加速其应用程序。

2.5 计算能力

计算能力 GP100、GV100 、GA100 对比如下：

Data center GPU	NVIDIA Tesla P100	NVIDIA Tesla V100	NVIDIA A100
GPU Codename	GP100	GV100	GA100
GPU Architecture	NVIDIA Pascal	NVIDIA Volta	NVIDIA Ampere
Compute Capability	6.0	7.0	8.0
Threads / Warp	32	32	32
Max Warps / SM	64	64	64
Max Threads / SM	2048	2048	2048
Max Thread Blocks / SM	32	32	32
Max 32-bit Registers / SM	65536	65536	65536
Max Registers / Block	65536	65536	65536
Max Registers / Thread	255	255	255
Max Thread Block Size	1024	1024	1024
FP32 Cores / SM	64	64	64
Ratio of SM Registers to FP32 Cores	1024	1024	1024
Shared Memory Size / SM	64 KB	Configurable up to 96 KB	Configurable up to 164 KB

图 2.5: 计算能力 GP100 vs . GV100 vs . GA100

3 总结

NVIDIA A100 Tensor Core GPU 在加速数据中心平台上实现了下一个巨大的飞跃，在各个规模上都提供了无与伦比的加速度。A100 支持许多应用领域，包括 HPC 、基因组学、5G、渲染、深度学习、数据分析、数据科学和机器人技术，有很好的应用前景

参考文献

[1] www.top500.org.

[2] https://baike.baidu.com/item/fromtitle=gpu&fromid=105524.

[3] https://www.nvidia.cn/data-center/a100/.

[4] https://baike.baidu.com/item/CUDA/1186262?fr=aladdin.