# Tesla GPU架构分析

在 2017 GPU 技术大会（GTC 2017）上，英伟达 CEO 黄仁勋正式发布了新一代处理器架构 Volta，以及使用新架构的第一款设备——适用于深度学习任务的加速卡 Tesla V100，英伟达将这块显卡称为全球最先进的数据中心 GPU。

从语音识别到训练虚拟助理进行自然交流，从探测车道线到让汽车完全自动驾驶，数据科学家们在技术的发展过程中正一步步攀登人工智能的新高度。而解决这些日益复杂的问题则需要日益复杂的深度学习模型，为深度学习提供强大的计算硬件是英伟达努力的目标。

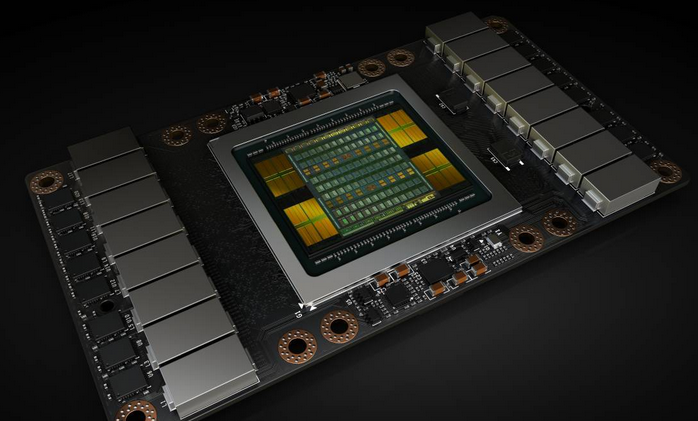


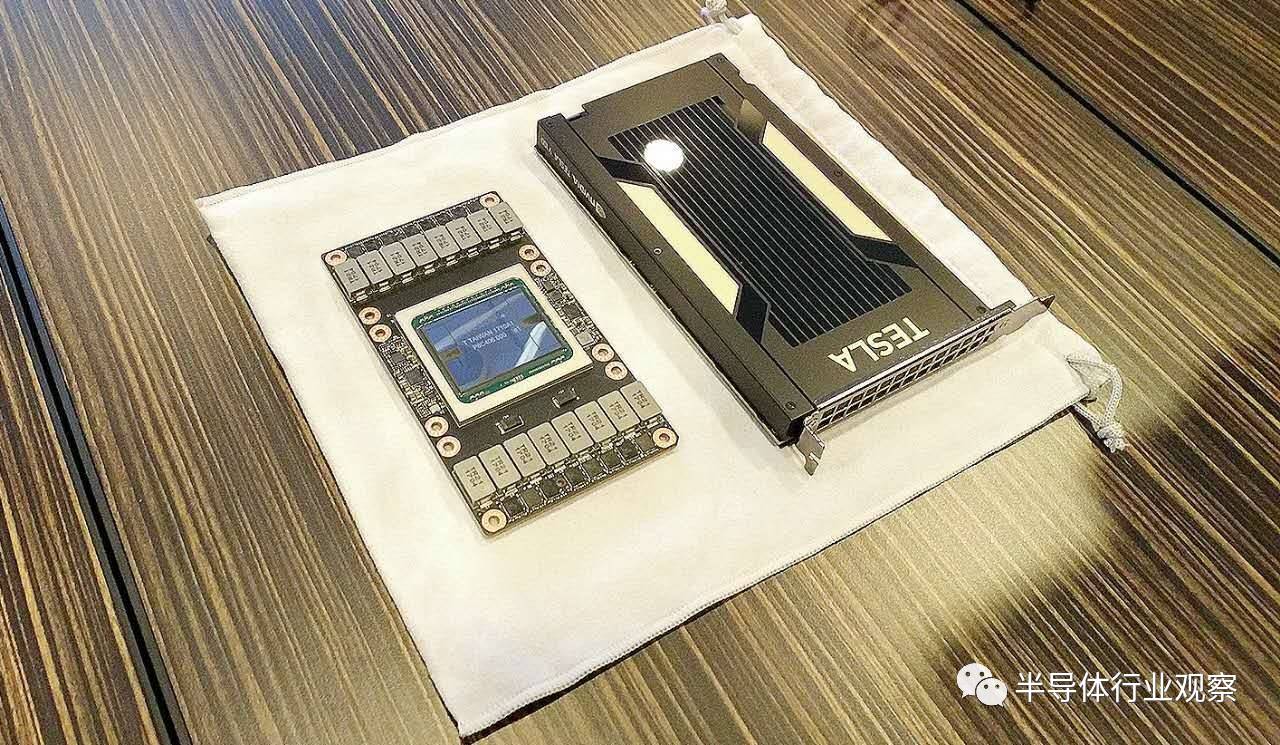
图 1． Tesla V100 加速卡内含 Volta GV100 GPU，以及 SXM2 Form Factor。

高性能计算设备（HPC）是现代科学的基础，从预测天气、发明新药到寻找新能源，大型计算系统能为我们模拟和预测世界的变化。这也是英伟达在新一代 GPU 架构推出时选择优先发布企业级计算卡的原因。黄仁勋在发布会上表示，全新的 Tesla V100 专为 HPC 和 AI 的融合而设计，同时采用了具有突破性的新技术。英伟达的新架构能否让 GPU 再上一个台阶？让我们随着 Tesla V100 一探究竟。

**揭秘新架构与 GPU 特性**

Volta 并不是 Pascal 的升级，而是一个全新的架构！——NVIDIA 应用深度学习研究副总裁 Bryan Catanzaro。

在 Nvdia GTC 2017 第三天下午，Nvidia CUDA 软件首席工程师 Luke Durant 与 Nvidia 首席构架师 Oliver Giroux 进行了一个名为 Inside Volta 的技术讲座，解读了 Volta 构架的设计。



英伟达认为，硬件的可编程性正在驱动深度学习的发展。谈到 Volta 对人工智能带来的影响时，英伟达副总裁 Bryan Catanzaro 表示，“Volta 提供大量的 FLOP，基于 Volta，人们就可以使用需要更多 FLOP 的模型。如今很多流行的模型都需要很大的计算资源，例如卷积，我个人认为架构上将会有一定的转向，既更多地来利用更多地利用我们已有的大量的 FLOP。当然，构架的进化也会经过一个‘达尔文’过程，最终最适应的会成为终极形态”

在本次 GTC 中，我们没有看到联网移动端芯片的身影或为移动端人工智能计算性能提升进行的构架设计，关于这个问题，Volta 设计团队表示，对于可以联网的设备，通过 CPU 结合 GPU 的混合云进行大量计算是必然趋势；而对于无法联网的应用场景，SOC 是更好的选择。

Tesla V100：人工智能计算和 HPC 的助推器

毫无疑问，全新的英伟达 Tesla V100 加速器是世界上性能最高的并行处理器，旨在为计算量最大的 HPC 设备、人工智能和图形工作任务提供支持。它的核心 GV100 GPU 包含 211 亿个晶体管，而芯片面积为前所未有的 815 平方毫米（Tesla GP100 为 610 平方毫米）。它采用了台积电（TSMC）的 12nm FFN 专属工艺打造。与其前身 GP100 GPU 及其他 Pascal 架构的显卡相比，GV100 提供了更强的计算性能，并增加了许多新功能。它进一步减小了 GPU 编程和应用程序移植难度，也通过制程的升级提高了 GPU 资源利用率。另外，GV 100 也是一款能效极高的处理器，其在单位功耗的性能上表现卓越。图 2 给出了 ResNet－50 深度神经网络在 Tesla V100 上进行训练的性能表现。

对于 12nm 制程的选择（AMD 准备在 2018 年推出使用 7nm 制程的显卡），英伟达的首席工程师表示他们已在功耗和性能之间做出了最佳选择。

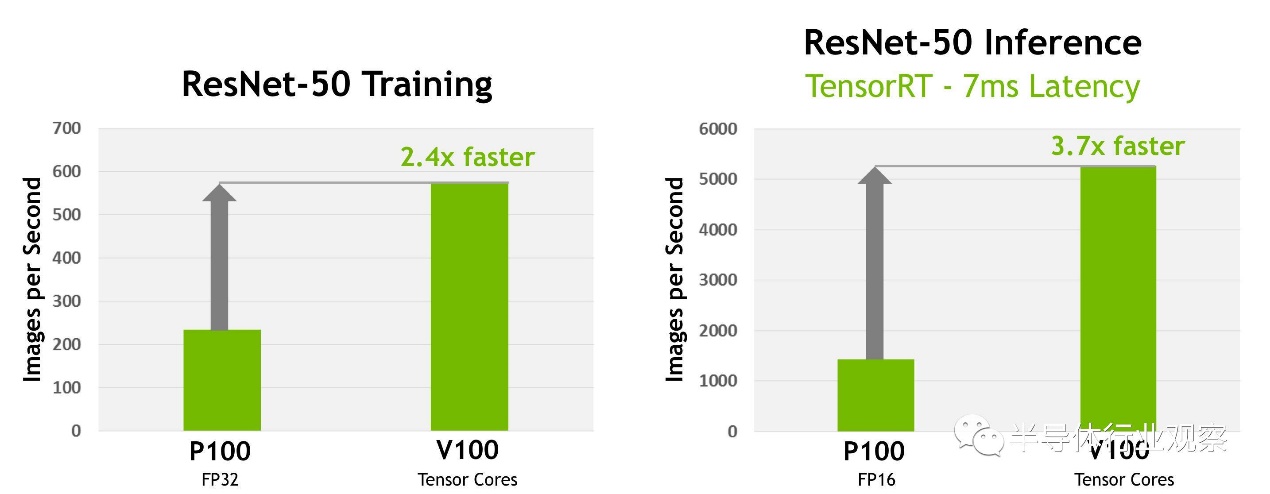


图 2． Tesla V100 在 ResNet－50 深度神经网络训练任务中的速度比 Tesla P100 快 2．4 倍。如果每张图像的目标延迟是 7ms，那么 Tesla V100 使用 ResNet－50 深度神经网络进行推理的速度比 P100 快 3．7 倍（参与测试的 V100 为原型卡）。

Tesla V100 的主要计算特征包括：

● 为深度学习优化过的新型流式多处理器（SM）架构。Volta 对 GPU 核心的 SM 处理器架构进行了重要的重新设计。新的 Volta SM 架构比前代 Pascal 设计能效高 50％，在同样的功率范围下 FP32 和 FP64 性能有重大提升。新的 Tensor Core 是专门为深度学习设计的，为浮点运算速度带来了 12 倍的提升。有了独立的、并行的整型和浮点型数据通路，Volta SM 在负载上也更高效，混合了计算与地址运算。Volta 新的独立线程调度能力使得并行线程之间的细粒度同步协同（finer－grain synchronization and cooperation）成为可能。最终，新型的 L1 Data Cache 与 Shared Memory 子系统的结合也能极大地提升性能，同时还简化了编程。

● 第二代 NVLink。第二代英伟达 NVLink 高速互连技术能提供更高的带宽、更多连接，同时还改进了多 GPU 和多 GPU／CPU 系统配置的延展性。

● HBM2 显存：更快、更高效。Volta 高度调整的 16GB HBM2 显存子系统提供了 900 GB／s 的峰值显存带宽。来自三星的新一代 HBM2 显存和 Volta 中的新一代显存控制器的组合实现的显存带宽是 Pascal GP100 的 1．5 倍，而且在许多负载上的显存带宽效率更高。

● Volta 多处理服务。Volta 多服务处理（MPS：Multi－Process Service）是 Volta GV100 的一项新特性，能够为 CUDA MPS 服务器的关键组件提供硬件加速，从而能为共享该 GPU 的多个计算应用提供更高的性能、隔离和更好的服务质量（QoS）。Volta MPS 还将 MPS 客户端的最大数量从 Pascal 的 16 提升到了 Volta 的 48。

● 增强统一存储和地址转换服务。Volta GV100 中的 GV100 统一存储（GV100 Unified Memory）技术包括新型访问计数器，让访问网页最频繁的处理器能更准确的迁移存储页。

● 协作组（Cooperative Groups）和新的 Cooperative Launch API。协作组是 CUDA 9 中新的编程模型，用来组织通信线程组。Volta 增加了对新型同步模式的支持。

● 最大性能和最大效率模式。在最大性能模式下，Tesla V100 加速器将不受限制的把 TDP（热设计功耗）水平提高到 300W，从而加速需要最快计算速度和最高数据吞吐的应用。最大效率模式下，数据中心管理员可以调整 Tesla V100 加速器的功率使用，从而用单位功耗下最优的性能进行运算。

● 为 Volta 优化过的软件。Caffe2、MXNet、CNTK、TensorFlow 等这样的深度学习框架的新版本，能够利用 Volta 的性能来获得更快的训练速度、更高的多节点训练性能。GPU 加速库（比如 cuDNN、cuBLAS 等）的 Volta 优化版本利用 Volta GV100 架构的新特性能为深度学习和高性能计算应用提供更高的性能。

GV100 GPU 硬件架构

装备有 Volta GV100 GPU 的英伟达 Tesla V100 加速器是目前世界上速度最快的并行计算处理器。GV100 的硬件创新十分显著，除了为 HPC 系统和应用提供远比现在更强的计算能力（如图 3 所示）之外，它还可以大大加快深度学习算法和框架的运行速度。

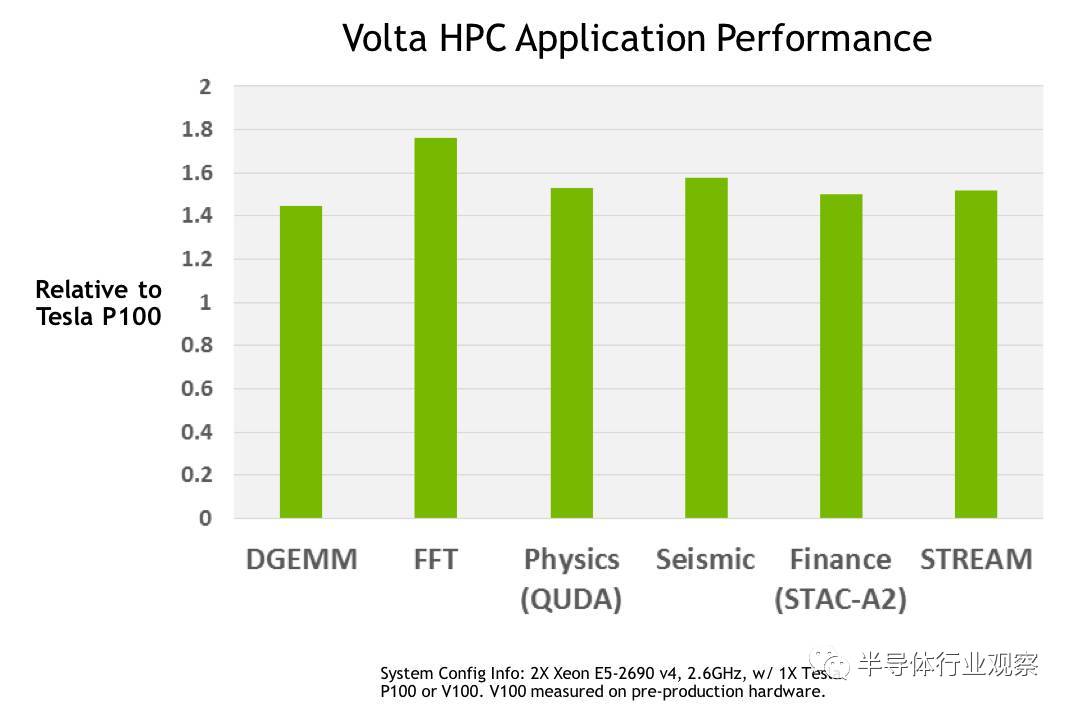


图 3：在各种 HPC 任务中，Tesla V100 平均比 Tesla P100 快 1．5 倍。（该性能基于 Tesla V100 原型卡）

Tesla V100 可以提供业界领先的浮点和整型计算性能。峰值计算速度（基于 GPU Boost 时钟频率）：

● 双精度浮点（FP64）运算性能：7．5 TFLOP／s；

● 单精度（FP32）运算性能：15 TFLOP／s；

● 混合精度矩阵乘法和累加：120 Tensor TFLOP／s。

与前一代 Pascal GP100 GPU 类似，GV100 GPU 由多个图形处理集群（Graphics Processing Cluster，GPC）、纹理处理集群（Texture Processing Cluster，TPC）、流式多处理器（Streaming Multiprocessor，SM）以及内存控制器组成。一个完整的 GV100 GPU 由 6 个 GPC、84 个 Volta SM、42 个 TPC（每个 TPC 包含了 2 个 SM）和 8 个 512 位的内存控制器（共 4096 位）。每个 SM 有 64 个 FP32 核、64 个 INT32 核、32 个 FP64 核与 8 个全新的 Tensor Core。同时，每个 SM 也包含了 4 个纹理处理单元。

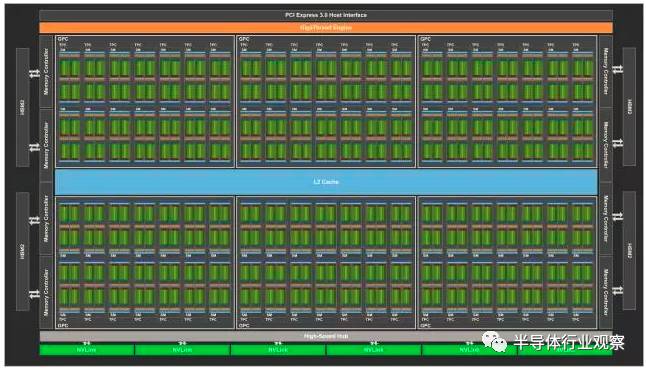


图 4：带有 84 个 SM 单元的完整 Volta GV100。

加上 84 个 SM，一个完整的 GV100 GPU 总共有 5376 个 FP32 核、5376 个 INT32 核、2688 个 FP64 核、672 个 Tensor Core 与 336 个纹理单元。每块内存控制器都连接了一个 768 KB 的 2 级缓存，每个 HBM2 DRAM 堆栈都由一对内存控制器控制。一个完整的 GV100 GPU 包括了总共 6144 KB 的二级缓存。图 4 展示了一个带有 84 个 SM 单元的完整 GV100 GPU（不同产品可以使用不同的 GV100 配置）。Tesla V100 加速器使用了 80 个 SM 单元。

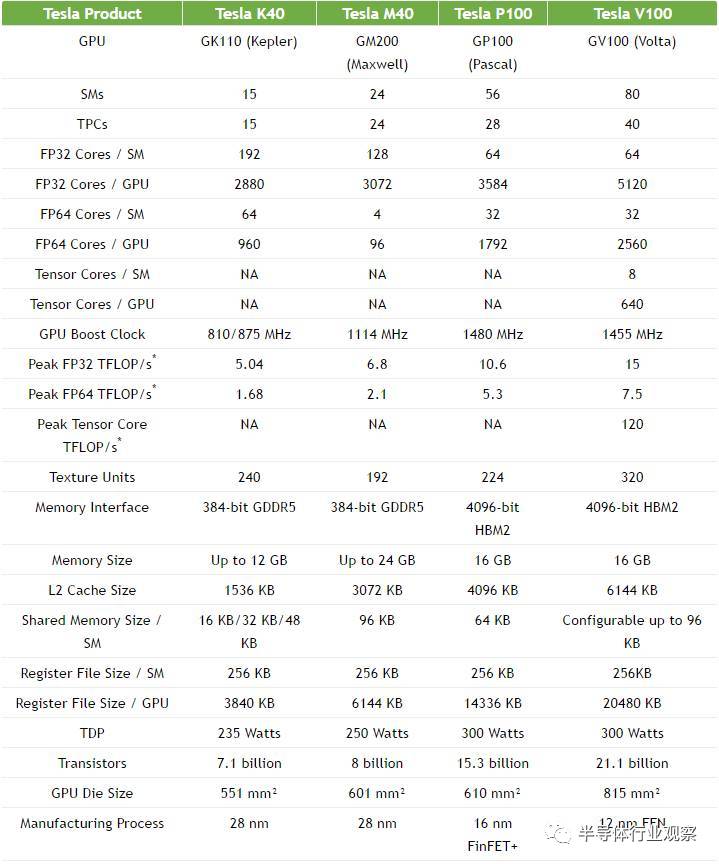


表 1． Tesla V100 与过去五年历代 Tesla 加速器的参数对比

Volta SM（流式多处理器）

为提供更高的性能而设计的架构，Volta SM 比过去的 SM 设计有更低的指令与缓存延迟，也包括加速深度学习应用的新特性。

主要特性包括：

● 为深度学习矩阵计算建立的新型混合精度 FP16／FP32 Tensor Core。

● 为更高的性能、更低的延迟而增强的 L1 数据缓存。

● 为更简单的解码而改进的指令集，并减少了指令延迟。

● 更高的速度和能效。

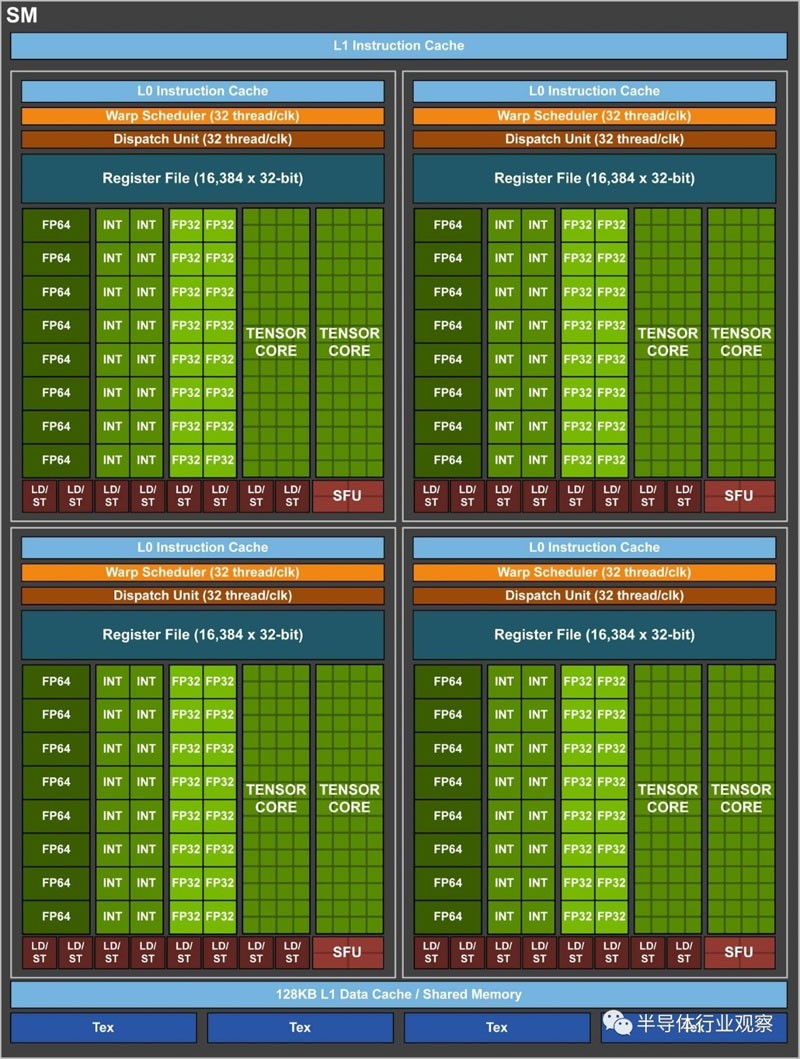


图 5： Volta GV100 SM

Tensor Core：深度学习专用核心

新的 Tensor Core 是 Volta GV100 最重要的特征，有助于提高训练神经网络所需的性能。Tesla V100 的 Tensor Core 能够为训练、推理应用的提供 120 Tensor TFLOPS。相比于在 P100 FP 32 上，在 Tesla V100 上进行深度学习训练有 12 倍的峰值 TFLOPS 提升。而在深度学习推理能力上，相比于 P100 FP16 运算，有了 6 倍的提升。Tesla V100 GPU 包含 640 个 Tensor Core：每个流式多处理器（SM）包含 8 个。

Tensor Core 非常省电，电力消耗大有可能将不再是深度学习的一大瓶颈。Bryan Catanzaro 表示：“通过底层数学计算的优化，Tensor Core 相较之前的构架要省电很多。深度学习的一个重要的限制是 energy efficiency，Tensor Core 在解决这个问题的方面相当突出。”

矩阵－矩阵乘法运算（BLAS GEMM）是神经网络训练和推理的核心，被用来获得输入数据和权重的大型矩阵的乘积。如下图 6 所示，相比于基于 Pascal 的 GP100，Tesla V100 中的 Tensor Core 把这些运算的性能提升了至少 9 倍。

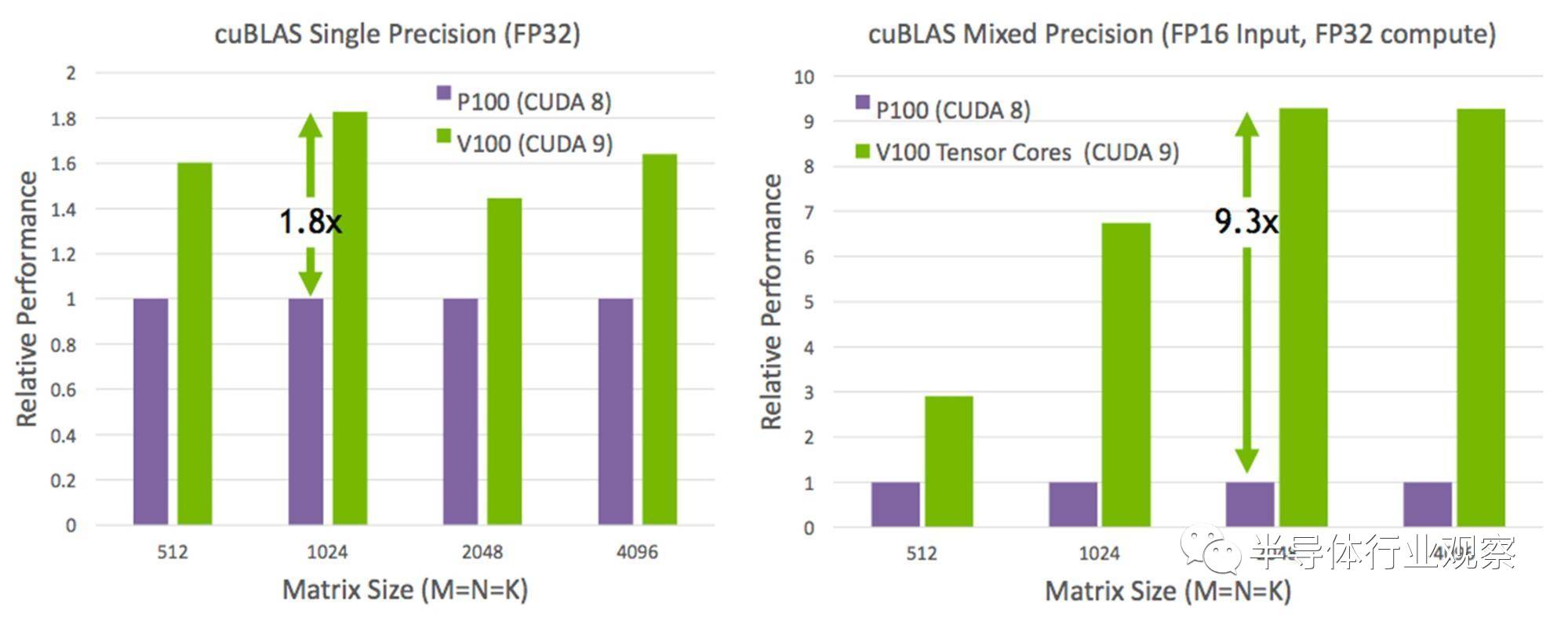


图 6：Tesla V100 Tensor Core 和 CUDA 9 对 GEMM 运算有了 9 倍的性能提升。（在 Tesla V100 样机上使用预发布的 CUDA 9 软件进行的测试）

Tensor Core 和与它们关联的数据通道进行了精心的定制，从而极大地提升了极小区域和能量成本下浮点计算的吞吐量。它也广泛地使用了时钟门控来尽可能节能。

每个 Tensor Core 包含一个 4x4x4 的矩阵处理阵列来完成 D＝A x B ＋ C 的运算，其中 A、B、C、D 是 4×4 的矩阵，如下图 7 中所示。矩阵相乘的输入 A 和 B 是 FP16 矩阵，相加矩阵 C 和 D 可能是 FP16 矩阵或 FP32 矩阵。

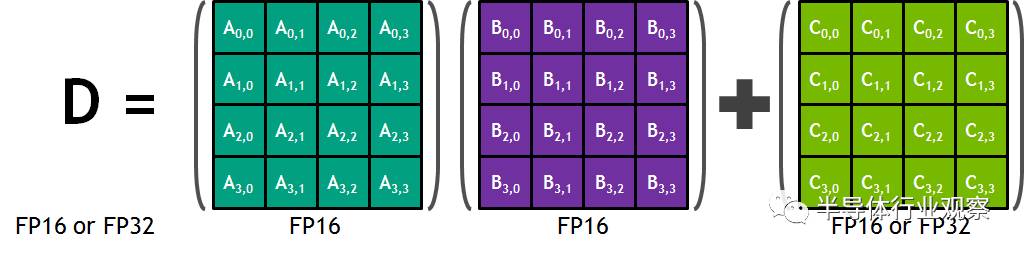


图 7：Tensor Core 的 4x4x4 矩阵乘法与累加。

每个 Tensor Core 每个时钟可执行 64 次浮点 FMA 混合精度运算（FP16 乘法与 FP32 累加），一个 SM 单元中的 8 个 Tensor Core 每个时钟可执行共计 1024 次浮点运算。相比于使用标准 FP32 计算的 Pascal GP100 而言，单个 SM 下的每个深度学习应用的吞吐量提升了 8 倍，所以这最终使得 Volta V100 GPU 相比于 Pascal P100 GPU 的吞吐量一共提升了 12 倍。Tensor Core 在与 FP32 累加结合后的 FP16 输入数据之上操作。FP16 的乘法得到了一个全精度结果，该结果在 FP32 和其他给定的 4x4x4 矩阵乘法点积的乘积运算之中进行累加。如图 8 所示。

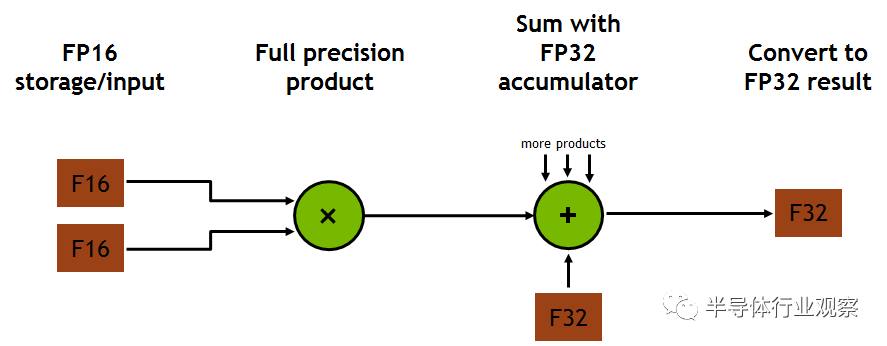


图 8． Volta GV100 Tensor Core 流程图

在程序执行期间，多个 Tensor Core 通过一组 warp 线程的执行而同时使用。warp 内的线程提供了 Tensor Core 来处理大型 16×16×16 矩阵运算。CUDA 将这些操作作为 Warp－Level 矩阵运算在 CUDA C＋＋ API 中公开。这些 C＋＋接口提供了专门化的矩阵负载，如矩阵乘法和累加，矩阵存储操作可以有效地利用 CUDA C＋＋程序中的 Tensor Core。

除 CUDA C＋＋接口可直接编程 Tensor Core 外，CUDA 9 cuBLAS 和 cuDNN 库还包含了使用 Tensor Core 开发深度学习应用和框架的新库接口。英伟达已经和许多流行的深度学习框架（如 Caffe2 和 MXNet）合作以使用 Tensor Core 在 Volta 架构的 GPU 系统上进行深度学习研究。英伟达将继续与其他框架开发人员合作以便在整个深度学习生态系统更广泛地使用 Tensor Core。

增强的 L1 数据缓存和共享显存

Volta SM 的 L1 数据缓存和共享显存子系统的组合能显著提高性能，同时也简化了编程并减少了达到或接近峰值应用性能所需的时间成本。

在共享显存块中进行集成可确保 Volta GV100 L1 缓存具有比过去英伟达 GPU 中的 L1 高速缓存更低的延迟和更高的带宽。L1 Volta 作为流式数据的高吞吐量导管（conduit），同时为经常复用的数据提供高带宽和低延迟访问，这两个性能都是目前最好的。英伟达表示，这一特性是 Volta 独有的，其提供比以往更强大的性能。

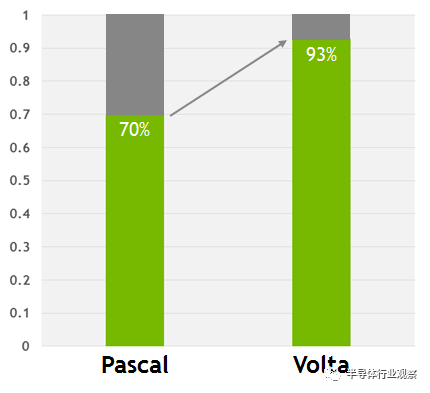


图 9． 在 Volta 上，这些代码在没有使用共享显存的情况下只有 7％的性能损失，而 Pascal 的性能下降了 30％。虽然共享显存仍然是最佳选择，但新 Volta L1 设计使程序员能够以更少的编程工作而快速获得足够出色的性能。

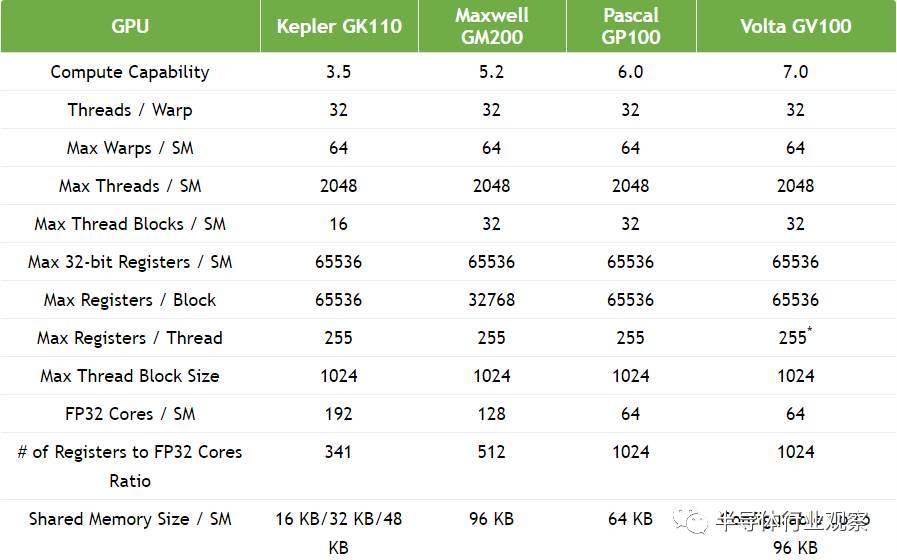


表 2． 全新 GV100 与此前各代架构 GPU 的形式比较，GV100 GPU 支持全新的 Compute Capability 7．0。

独立的线程调配

Volta 架构旨在设计为比以前的 GPU 更容易编程，令用户能在更复杂和多样的应用程序上高效地工作。Volta GV100 是第一款支持独立线程调配的 GPU，其在并行线程指令中可以实现细粒度（finer－grain）的同步和协作。Volta 主要的设计目标是减少在 GPU 中运行指令所需的工作量，并在线程合作中实现更大的灵活度，这样从而为细粒度并行算法提供更高的效率。

英伟达同时也展示了他们如何对 SIMT（单指令多线程）做出重大改进以推进 Volta 架构。32 线程内单个独立的 CUDA 核现在只有有限的自主性；线程现在可以在一个细粒度层面上进行同步，并且仍然处于 SIMT 范式下，所以这就意味着更高的整体效率。更重要的是，独立的线程现在可以放弃再重新安排在一起。这就意味着英伟达的 GPU 有一定数量的调度硬件（scheduling hardware）。

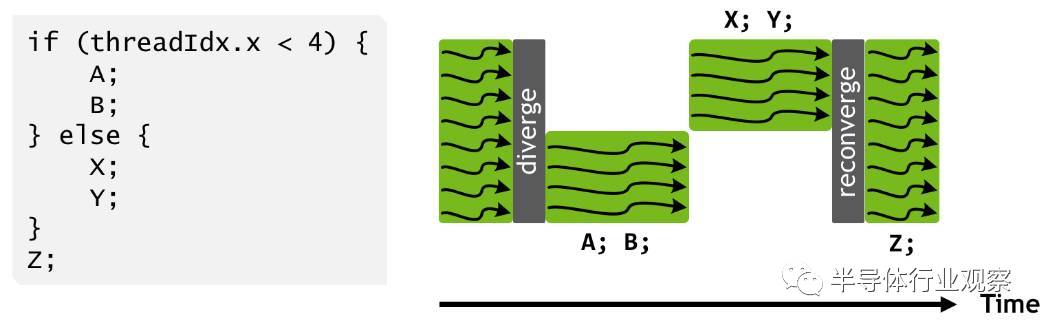


图 10：Pascal 和早期的英伟达 GPU 在 SIMT warp 执行模式下的线程调配。大写字母代表指令伪代码中的语句。在一个 warp 中不同的分支是序列化的，这样在分支一边的所有语句一起执行以在另一边语句执行前完成。在 else 语句后，warp 中的线程通常会重新映射。

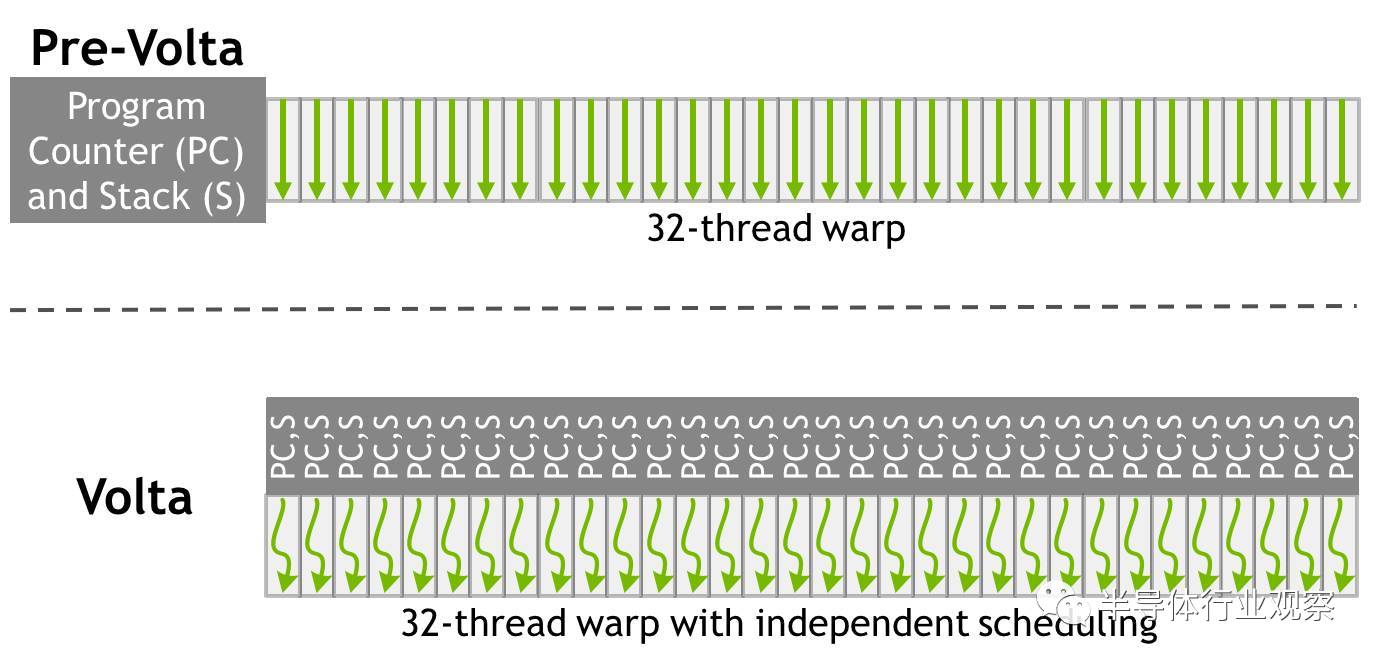


图 11：Volta（下方）独立线程调配架构图与 Pascal 和早期的架构（上方）相比较。Volta 会维持每个线程调配的资源，就像程序计数器（PC）和调用堆栈（S）那样，而早期的架构以每个 warp 为单位维持。

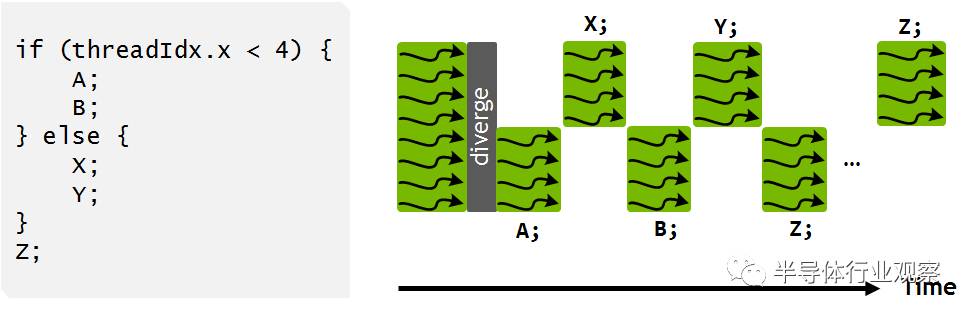


图 12：Volta 独立线程调配令其可以交叉执行发散分支（divergent branches）的语句。这就允许执行细粒度并行算法，而 warp 内的线程也就能同步和通信。

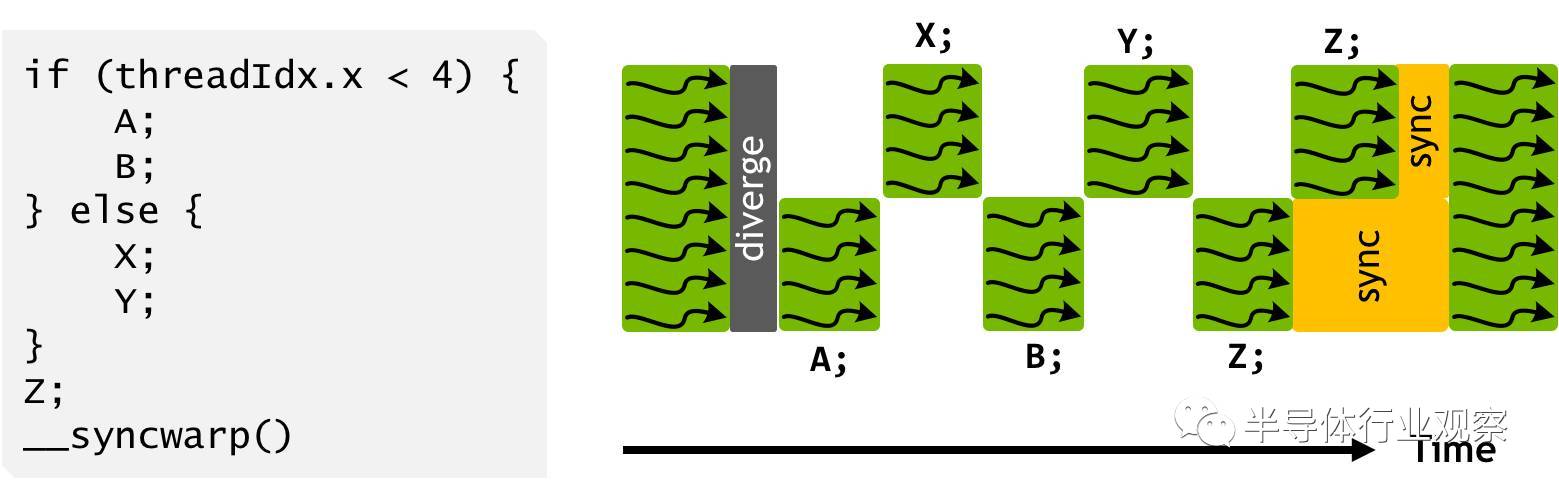


图 13：程序可以使用显式同步来重新令 warp 中的线程收敛

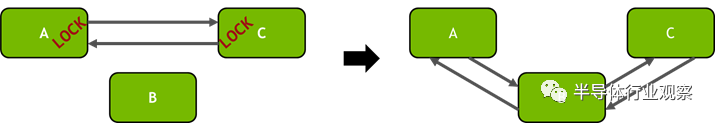


图 14：具有细粒度锁（fine－grained locks）的双向链表。在插入结点 B 到链表（右）前，需要获取每一个结点锁（左）

总结

根据目前公布的数字，Tesla V100 可以提供 15 TFLOPS 的 FP32、30 TFLOPS FP16、7．5 TFLOPS FP64 和高达 120 TFLOPS 的专用 Tensor 运算性能。由于 1455 MHz 的峰值运算速度，它相比前一代的 CUDA 理论 FLOPS 数据增长了 42％。Tesla V100 配备了 16G 的 HBM2 显存，它的内存时钟速度从 1．4 Gbps 提升至 1．75 Gbps，提升了 25％。