**Tesla GPU架构分析**

**目录:**

**Tesla V100 GPU**

**Tesla P100 GPU**

**一、Tesla V100 GPU**

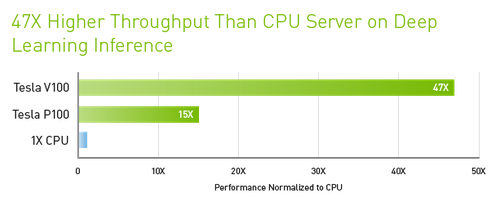
NVIDIA Tesla V100 是当今市场上为加速人工智能、高性能计算和图形的数据中心 GPU 中的精尖之作。NVIDIA Tesla V100 加速器基于全新 Volta GV100 GPU，Volta 是全球功能强大无比的 GPU 架构，而 GV100 是第一种突破 100 TFLOPS 深度学习性能极限的处理器。GV100 将 CUDA 核心和 Tensor 核心相结合，在 GPU 中提供 AI 超级计算机的出色性能。

NVIDIA Tesla V100 Tensor Core 是有史以来极其先进的数据中心 GPU，能加快 AI、高性能计算 (HPC) 和图形技术的发展。其采用 NVIDIA Volta 架构，并带有 16 GB 和 32GB 两种配置，在单个 GPU 中即可提供高达 100 个 CPU 的性能。

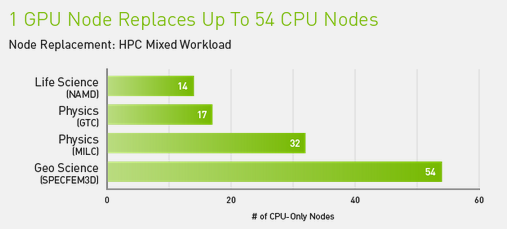
Tesla V100 拥有 640 个 [Tensor 内核](https://www.nvidia.cn/data-center/tensorcore/)，是世界上第一个突破 100 万亿次 (TFLOPS) 深度学习性能障碍的 GPU。新一代 NVIDIA NVLink 以高达 300 GB/s 的速度连接多个 V100 GPU，在全球打造出功能极其强大的计算服务器。现在，在之前的系统中需要消耗数周计算资源的人工智能模型在几天内就可以完成训练。随着训练时间的大幅缩短，人工智能现在可以解决各类新型问题。



由于将人工智能作为核心，Tesla V100 GPU 可提供比 CPU 服务器高 30 倍的推理性能。这种吞吐量和效率的大幅提升将使人工智能服务的扩展变成现实。

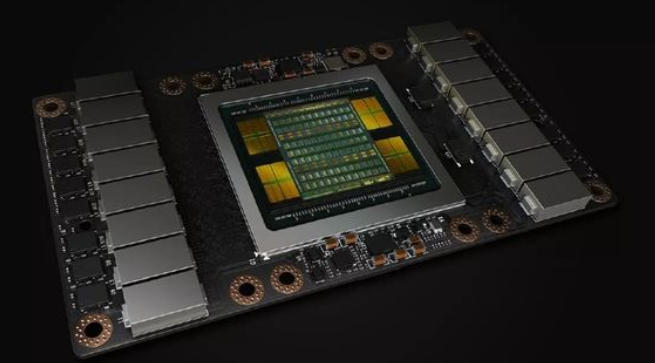


Tesla V100 的设计能够融合人工智能和高性能计算。它为高性能计算系统提供了一个平台，在用于科学模拟的计算机科学和用于在数据中发现见解的数据科学方面表现优异。通过在一个统一架构内搭配使用 NVIDIA CUDA内核和[Tensor 内核](https://www.nvidia.cn/data-center/tensorcore/)，配备 Tesla V100 GPU 的单台服务器可以取代数百台仅配备通用 CPU 的服务器来处理传统的高性能计算和人工智能工作负载。现在，每位研究人员和工程师都可以负担得起使用人工智能超级计算机处理最具挑战性工作的做法。



**基于 NVIDIA Volta 架构的 GV100 GPU**

Tesla V100 采用与 Tesla P100 相同的 SXM2 主板外形，大小为 140×78 毫米，主要区别在于 GPU 由 GV100 代替了 GP100。SXM2 主板支持 NVLink 和 PCIe 3.0 连接功能，包含可为 GPU 供应各种所需电压的高效电压调节器，额定为 300 瓦热设计功耗（TDP）。工作站、服务器和大型计算系统中可应用一个或多个 Tesla V100 加速器。



NVIDIA Tesla V100 加速器配备 Volta GV100 GPU，是世界领先的高性能并行处理器。包含 211 亿个晶体管，采用台积电 12nm FFN 专属工艺打造的它芯片面积达到了前所未有的 815 平方毫米（Tesla GP100 为 610 平方毫米）。

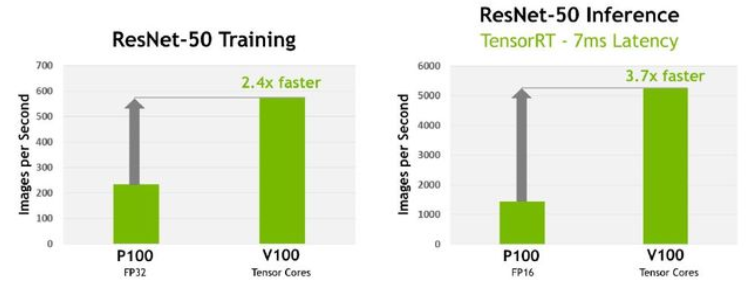
与上一代 Pascal GP100 GPU 一样，GV100 GPU 由 6 个 GPU 处理集群（GPC）和 8 个 512 位内存控制器组成，每个 GPC 拥有 7 个纹理处理集群（TPC），每个 TPC 含 2 个流多处理器（SM）。

含 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。



(含84个SM单元的完整Volta GV100 GPU)

**与其前身 GP100 GPU 及其他 Pascal 架构的显卡相比，GV100 在人工智能方面有非常出色的表现。**

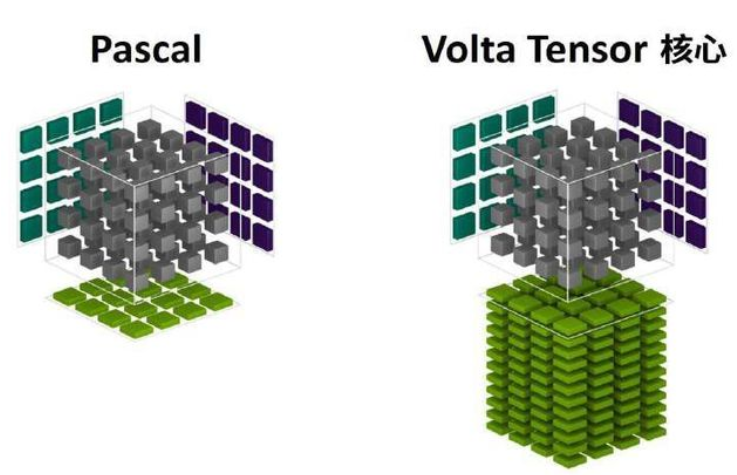


以 ResNet-50 深度神经网络为例，Tesla V100 的深度神经网络训练任务中的速度比 Tesla P100 快 2.4 倍。如果每张图像的目标延迟是 7ms，那么 Tesla V100 使用 ResNet-50 深度神经网络进行推理的速度比 P100 快 3.7 倍。

**强大的硬件规格也让 Tesla V100 具备了业界领先的浮点和整数性能。**

计算速率峰值在不同条件下，分别可以实现（基于 GPU Boost 时钟频率）：  
双精度浮点（FP64）运算性能：7.8 TFLOP/s 的双精度浮点：  
15.7 TFLOP/s 的单精度（FP32）性能   
125 Tensor TFLOP/s  


**Tensor Core：处理神经网络的最佳选择**

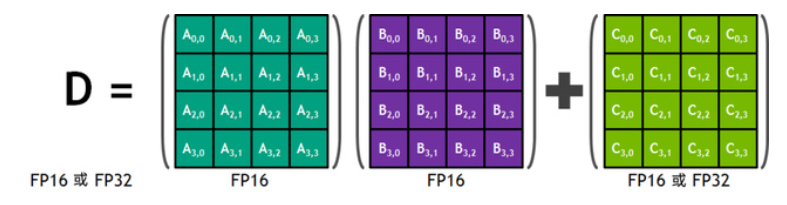


新的 Tensor Core 是 Volta GV100 最重要的特征，是其提供大型神经网络训练所需性能的关键。

Tesla V100 GPU 包含 640 个 Tensor 核心：每个 SM 有 8 个核心，SM 内的每个处理块（分区）有 2 个核心。在 Volta GV100 中，每个 Tensor 核心每时钟执行 64 次浮点 FMA 运算，一个 SM 中的 8 个 Tensor 核心每时钟总共执行 512 次 FMA 运算（或 1024 次单个浮点运算）。

Tesla V100 的 Tensor Core 能够为训练、推理的应用提供高达 125 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。

以上出色的表现实际上源于 Tensor Core 的基础架构，它针对人工智能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。



**新一代 NVLINK**

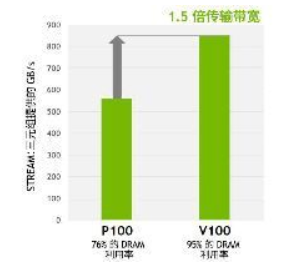
Tesla V100 中采用的 NVIDIA NVLink 可提供 2 倍于上一代的吞吐量。8 块 Tesla V100 加速器能以高达 300 GB/s 的速度互联，从而发挥出单个服务器所能提供的最高应用性能。

除了速度大幅提升之外，第二代 NVLink 还允许 CPU 对每个 GPU 的 HBM2 内存进行直接加载/存储/原子访问。NVLink 还支持一致性运算，允许读取自图形内存的数据存储在 CPU 的缓存层次结构中，进一步释放 CPU 性能。



**HBM2 内存架构**

相比上一代 Tesla P100，Tesla V100 采用了更快、更高效的 HBM2 架构。四个 HBM 芯片（堆栈）总共可以提供 900 GB/s 峰值内存带宽（上一代为 732GB/s）。同时 Volta 还采用了全新的内容控制器，也让内存带宽方面的优势进一步放大。在 STREAM 上测量时可提供高于 Pascal GPU 1.5 倍的显存带宽。



**二、Tesla P100 GPU**

NVIDIA TeslaP100 运用 [NVIDIA Pascal GPU 架构](https://www.nvidia.com/zh-cn/data-center/pascal-gpu-architecture/)提供统一的平台，以加速 HPC 和 AI 发展，大幅提升吞吐量，同时降低成本。

**特性和优势**

**1.Pascal 架构带来巨大的性能飞跃**

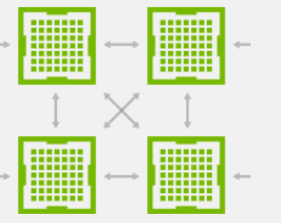
NVIDIA Pascal 架构使 Tesla P100 能为 HPC 和超大规模工作负载提供卓越性能。凭借每秒超过 21 万亿次的 16 位浮点 (FP16) 运算性能，经过优化的 Pascal 为[深度学习应用程序](https://www.nvidia.com/zh-cn/deep-learning-ai/)带来了令人兴奋的新可能。Pascal 还可为 HPC 工作负载提供超过 5 万亿次的双精度浮点运算和 10 万亿次的单精度浮点运算能力。

**2.通过采用 HBM2 的 CoWoS 技术实现更高效率**

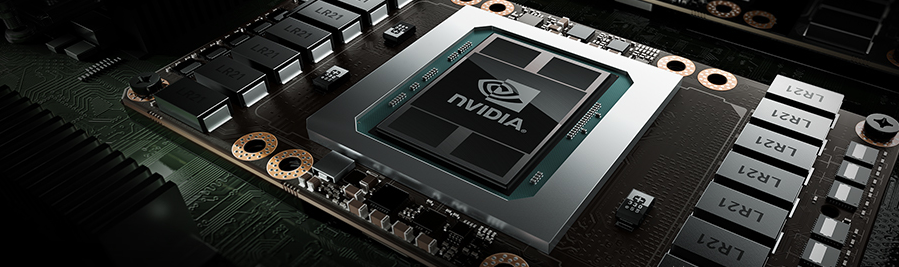
通过加入采用 HBM2 的晶圆基底芯片 (CoWoS) 技术，Tesla P100 将计算性能和数据紧密集成在同一个程序包内，提供的内存性能是 [NVIDIA Maxwell 架构的](https://developer.nvidia.com/maxwell-compute-architecture)三倍以上。这大幅缩短了为数据密集型应用程序算出解决方案的时间。

**3**. **NVIDIA NVLink 让超大规模的应用成为现实**

互连技术通常制约着性能。创新的 [NVIDIA NVLink 高速双向互连](https://www.nvidia.com/zh-cn/data-center/nvlink/)技术能跨越多个 GPU 扩展应用程序，其性能比当今的一流技术高 5 倍



Tesla P100 采用顶级大核心 GP100，面积 610 mm^2。



GP100 参数汇总如下：

芯片：GP100

架构：SM\_60

工艺：16 nm FinFET

支持：双精度 FP64, 单精度 FP32, 半精度 FP16

功耗：250 W

CUDA 核心数：3584（56 SMs, 64 SPs/SM)

GPU 时钟（Base/Boost)：1189 MHz/1328 MHz

PCIe：Gen 3 x16

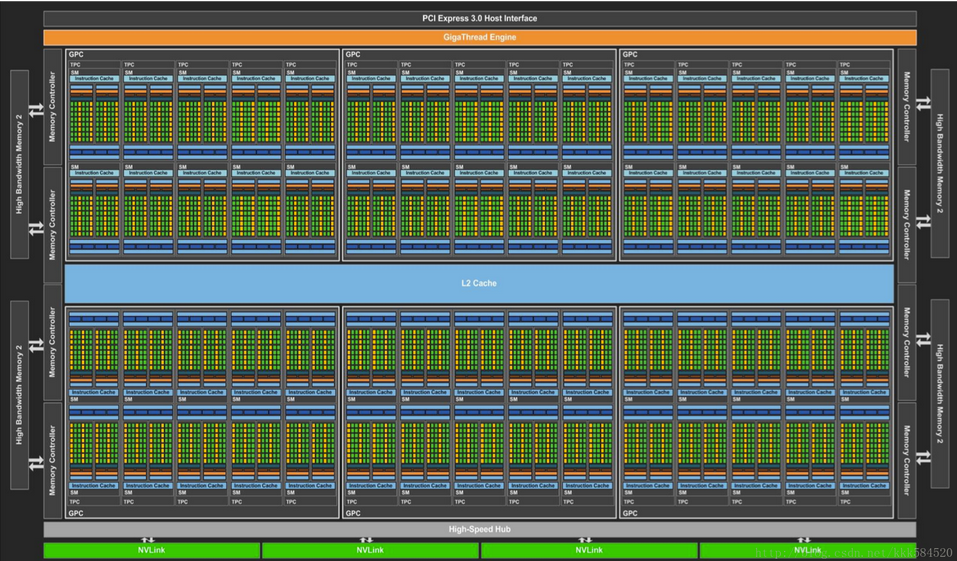
显存容量：12/16 GB HBM2

显存位宽：3072/4096 bits

显存时钟：715 MHz

显存带宽：539/732 GB/s

GP100 包含一组 GPC（图形处理簇，Graphics Processing Clusters）、TPC（纹理处理簇，Texture Processing Clusters）、SM（流多处理器，Stream Multiprocessors）以及内存控制器。其架构如下：



一颗完整的 GP100 芯片包括 6 个图形处理簇，60 个 Pascal 流多处理器，30 个纹理处理簇和 8 个 512 位内存控制器（总共 4096 位）。

每个图形处理簇内部包括 10 个 流多处理器。

每个流多处理器内部包括 64 个 CUDA 核心和 4 个纹理单元。

GP100 的 SM 包括 64 个单精度 CUDA 核心。而 Maxwell 和 Kepler 的 SM 分别有 128 和 192 个单精度 CUDA 核心。虽然 GP100 SM 只有 Maxwell SM 中 CUDA 核心数的一半，但总的 SM 数目增加了，每个 SM 保持与上一代相同的寄存器组，则总的寄存器数目增加了。这意味着 GP100 上的线程可以使用更多寄存器，也意味着 GP100 相比旧的架构支持更多线程、warp 和线程块数目。与此同时，GP100 总共享内存量也随 SM 数目增加而增加了，带宽显著提升不至两倍。



图中为一个 SM 的架构。其中绿色的“Core” 为单精度 CUDA 核心，共有 64 个，同时支持 32 位单精度浮点计算和 16 位半精度浮点计算，其中 16 位计算吞吐是 32 位计算吞吐的两倍。图中橘黄色的 “DP Unit” 为双精度计算单元，支持 64 位双精度浮点计算，数量为 32 个。每个 GP100 SM 双精度计算吞吐为单精度的一半。