TESLA V100
性能指南

深度学习和 HPC 应用程序

NVIDIA

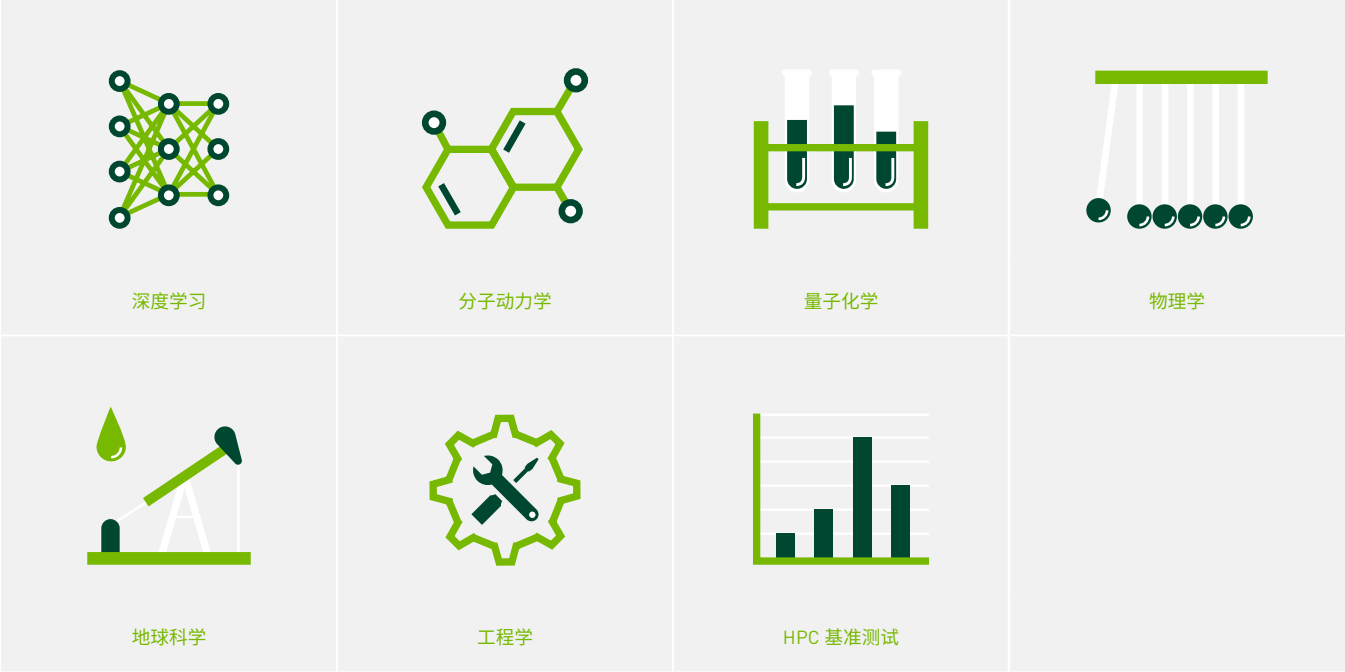
2017 年 11 月

TESLA V100 性能指南

现代高性能计算（HPC）数据中心是解决全球一些重大科学和工程挑战的关键。NVIDIA® Tesla® 加速计算平台让这些现代数据中心能够使用行业领先的应用程序加速完成 HPC 和 AI 领域的工作。Tesla V100 GPU 是现代数据中心的引擎，能够以更少的服务器提供突破性性能，从而加快探索发现的步伐，并大幅降低成本。改进的性能和解决方案时间对提高收益和生产力也有显著的有利影响。

每个 HPC 数据中心都可从 Tesla 平台中受益。多个领域超过 500 款 HPC 应用程序已经过 GPU 优化，其中包括全部 15 大常用 HPC 应用程序以及各主要的深度学习框架。

使用 GPU 加速应用程序的研究领域包括：

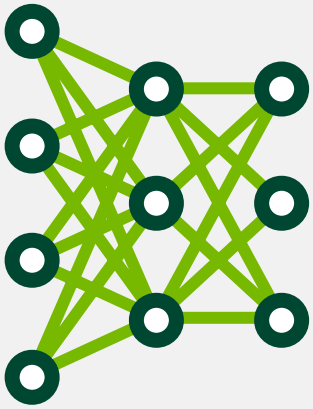


超过 500 款 HPC 应用程序和各主要深度学习框架已支持 GPU 加速。

- 要获取 GPU 加速应用程序的最新目录，请访问：
  www.nvidia.cn/object/gpu-applications-cn
- 如要获取适用于各种加速应用程序的简单指令，以实现 GPU 上的快速启动和运行，请访问：
  www.nvidia.com/gpu-ready-apps

TESLA V100 性能指南

深度学习



深度学习正在解决几年前还看似遥不可及的科学、企业级的和消费者层面的重要问题。各主要深度学习框架均支持 NVIDIA GPU 优化，从而使数据科学家和研究人员可以在工作中利用人工智能。数据中心配有 Tesla V100 GPU 后，在运行深度学习训练和推理框架时可节省高达 85% 的服务器和基础架构购置成本。

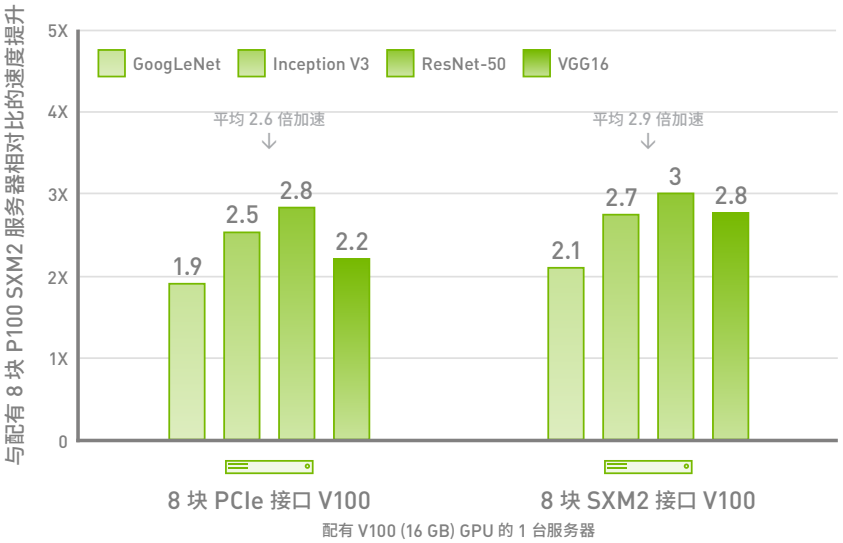
深度学习训练适用的 TESLA 平台和 V100 的主要特性

- 借助 Tesla V100，Caffe、TensorFlow 和 CNTK 的速度可提升至高达 3 倍（与 P100 相比）
- 所有常用深度学习框架均支持 GPU 加速
- TensorFlow 运算能力高达 125 TFLOPS/s
- 显存容量高达 16 GB，显存带宽高达 900 GB/s

如要查看所有相关应用程序，请访问：
www.nvidia.cn/object/gpu-applications-cn

Caffe 深度学习框架

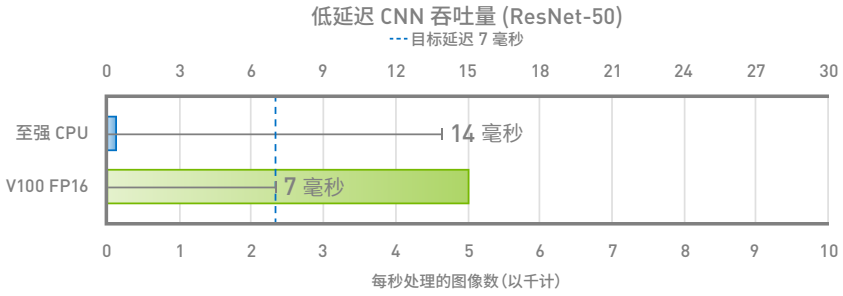
配有 8 块 V100 GPU 的服务器与 8 块 P100 GPU 服务器上的训练对比



CPU 服务器：双路至强 E5-2698 v4 @ 3.6GHz，GPU 服务器如图所示 | Ubuntu 14.04.5 | CUDA 版本：CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | 驱动程序 384.66 | 数据集：ImageNet | 批量大小：GoogleNet 192、Inception V3 96、ResNet-50 64（用于 P100 SXM2）和 128（用于 Tesla P100）、VGG16 96

低延迟 CNN 推理性能

巨大的吞吐量和惊人的低延时效率



系统配置：至强 E2690 v4 @ 3.5GHz，单卡 NVIDIA® Tesla® V100，运行 TensorRT 3 RC 与 Intel DL SDK beta 2 的 GPU | Ubuntu 14.04.5 | CUDA 版本：7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | 驱动程序 384.66 | 精度：CPU FP32，NVIDIA Tesla V100 FP16

CAFFE

加州大学伯克利分校开发的一种热门 GPU 加速深度学习框架

版本

1.0

加速特性

全框架加速

可扩展性

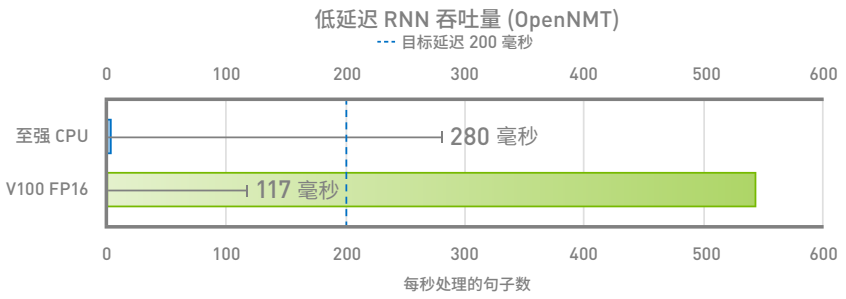
多 GPU

更多信息

caffe.berkeleyvision.org

低延迟 RNN 推理性能

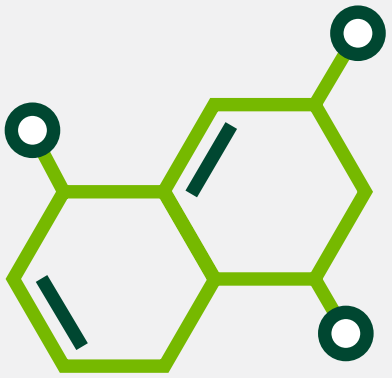
巨大的吞吐量和惊人的低延时效率



系统配置：至强 E2690 v4 @ 3.5GHz，单卡 NVIDIA® Tesla® V100，运行 TensorRT 3 RC 与 Intel DL SDK beta 2 的 GPU | Ubuntu 14.04.5 | CUDA 版本：7.0.1.13 | CUDA 9.0.176 | NCCL 2.0.5 | CuDNN 7.0.2.43 | 驱动程序 384.66 | 精度：CPU FP32，NVIDIA Tesla V100 FP16

TESLA V100 性能指南

分子动力学



分子动力学 (MD) 代表 HPC 数据中心的大部分工作负载。所有常用 MD 应用程序均已支持 GPU 加速，科学家们先前无法借助这些应用程序的传统纯 CPU 版本执行的模拟，现在都可运行。数据中心配有 Tesla V100 GPU 后，在运行 MD 应用程序时可节省高达 80% 的服务器和基础架构购买成本。

MD 适用的 TESLA 平台和 V100 的主要特性

- 对于 HOOMD-Blue 和 Amber 等应用程序，配有 V100 的服务器可以代替 54 台 CPU 服务器的对应性能
- 所有常用 MD 应用程序均支持 GPU 加速
- 支持主要数学库，例如 FFT 和 BLAS
- 每个 GPU 的单精度浮点运算能力高达 15.7 TFLOPS/s
- 每个 GPU 的带宽高达 900 GB/s

如要查看所有相关应用程序，请访问：
www.nvidia.cn/object/molecular_dynamics_cn

HOOMD-Blue 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.145 | 数据集：Microsphere | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

HOOMD-BLUE

专为 GPU 编写的粒子动力学程序包

版本

2.1.6

加速特性

CPU 和 GPU 版本可用

可扩展性

多 GPU 和多节点

更多信息

http://codeblue.umich.edu/hoomd-blue/index.html

AMBER 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：PME-Cellulose_NVE | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

AMBER

一套模拟生物分子层面上分子动力学的程序

版本

16.8

加速特性

PMEMD 显式溶剂和 GB；显式和隐式溶剂、REMD、aMD

可扩展性

多 GPU 和单节点

更多信息

http://ambermd.org/gpus

TESLA V100 性能指南

量子化学

量子化学 (QC) 模拟是发现新药物和材料的关键，占用 HPC 数据中心工作负载的大部分。目前，60% 的常用 QC 应用程序已支持 GPU 加速。数据中心配有 Tesla V100 GPU 后，在运行 QC 应用程序时可节省高达 30% 的服务器和基础架构购买成本。

QC 适用的 TESLA 平台和 V100 的主要特性

- 对于 VASP 等应用程序，配有 V100 的服务器可以代替多达 5 台 CPU 服务器的对应性能
- 60% 的常用 QC 应用程序均支持 GPU 加速
- 支持主要数学库，例如 FFT 和 BLAS
- 每个 GPU 的双精度浮点运算能力高达 7.8 TFLOPS/s
- 适用于大型数据集的显存容量高达 16 GB

如要查看所有相关应用程序，请访问：
www.nvidia.cn/object/computational_chemistry_cn

VASP 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

配有 V100 GPU 的 1 台服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：Si-Huge | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

VASP

用于执行量子力学分子动力学 (MD) 从头计算模拟的程序包

版本

5.4.4

加速特性

RMM-DIIS、Blocked Davidson、K-points 和精确交换

可扩展性

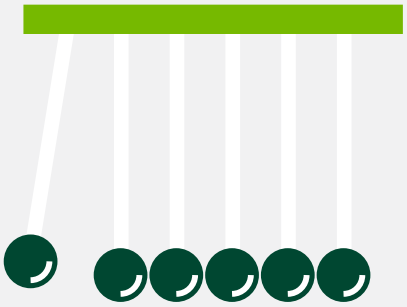
多 GPU 和多节点

更多信息

www.nvidia.com/vasp

TESLA V100 性能指南

物理学



从聚变能到高能粒子，物理学模拟涵盖了 HPC 数据中心的各种应用程序。许多常用物理学应用程序均已支持 GPU 加速，取得了之前不可能得到的宝贵见解。数据中心配有 Tesla V100 GPU 后，运行 GPU 加速的物理学应用程序时可节省高达 75% 的服务器购买成本。

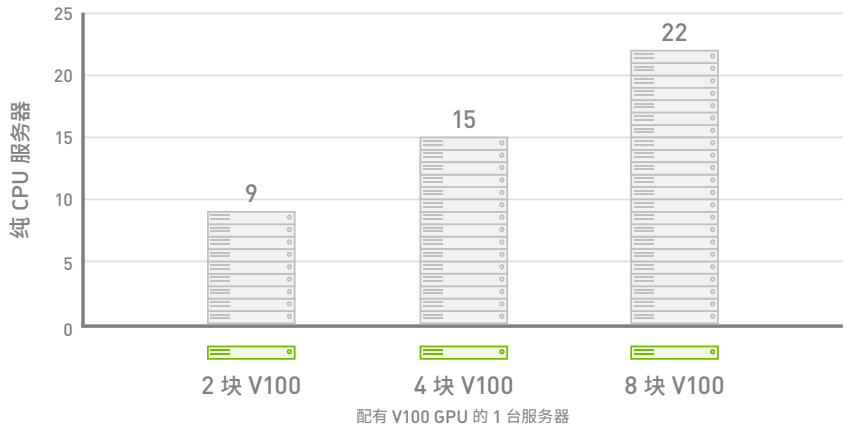
物理学适用的 TESLA 平台和 V100 的主要特性

- 对于 GTC-P、QUDA 和 MILC 等应用程序，配有 V100 的服务器可以代替多达 75 台 CPU 服务器的对应性能
- 大多数常用物理学应用程序均支持 GPU 加速
- 双精度浮点运算能力高达 7.8 TFLOPS/s
- 显存容量高达 16 GB，显存带宽高达 900 GB/s

如要查看所有相关应用程序，请访问：
www.nvidia.cn/object/gpu-applications-cn
（点击下拉菜单选择 “物理学”）

GTC-P 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：A.txt | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

GTC-P

用于优化等离子体物理学的开发代码

版本

2017

加速特性

推动、移动和碰撞

可扩展性

多 GPU

更多信息

www.nvidia.com/gtc-p

QUDA 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：Dslash Wilson-Clove；精度：单；Gauge Compression/Recon：12；问题大小：32x32x32x64 | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

QUDA

用于格点量子色动力学的 GPU 库

版本

2017

加速特性

全部

可扩展性

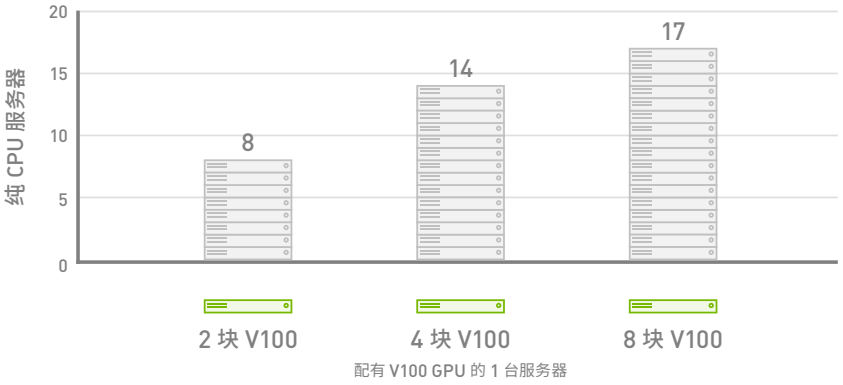
多 GPU 和多节点

更多信息

www.nvidia.com/quda

MILC 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：Precision=FP64 | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

MILC

格点量子色动力学 (LQCD) 代码，用于模拟基本粒子如何通过“强相互作用”形成和束缚以生成更大的粒子（例如质子和中子）

版本

2017

加速特性

交错费米子、克里洛夫求解和链节增大

可扩展性

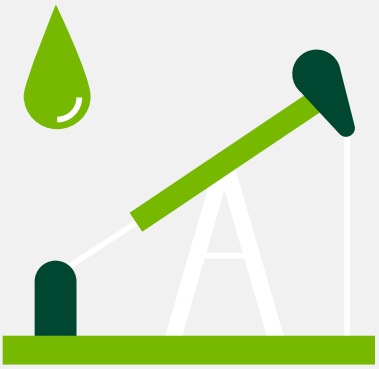
多 GPU 和多节点

更多信息

www.nvidia.com/milc

TESLA V100 性能指南

地球科学



地质科学模拟是发现石油和天然气以及执行地质建模的关键。目前，许多常用地质科学应用程序均已支持 GPU 加速。数据中心配有 Tesla V100 GPU 后，在运行地球科学应用程序时可节省高达 70% 的服务器和基础架构购买成本。

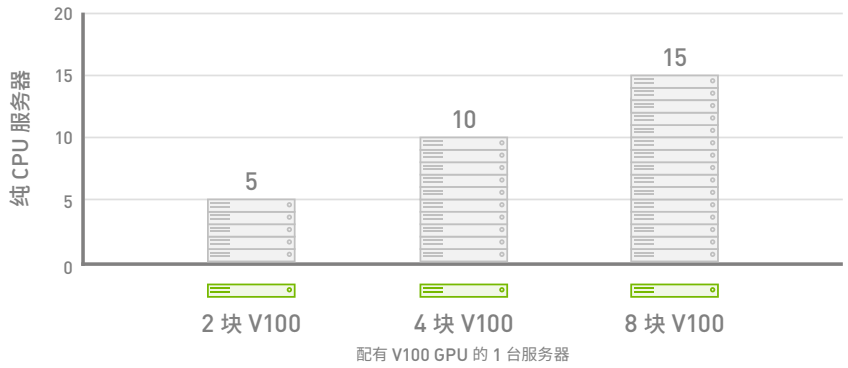
地球科学适用的 TESLA 平台和 V100 的主要特性

- 对于 RTM 和 SPECFEM 3D 等应用程序，配有 V100 的服务器可以代替多达 82 台 CPU 服务器的对应性能
- 常用石油和天然气应用程序均支持 GPU 加速
- 单精度浮点运算能力高达 15.7 TFLOPS/s
- 显存容量高达 16 GB，显存带宽高达 900 GB/s

如要查看所有相关应用程序，请访问：
www.nvidia.cn/object/gpu-applications-cn
（点击下拉菜单选择 “石油天然气 / 地震”）

RTM 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：TTI RX 2pass mgpu | 为了达到与 CPU 节点相同的性能，我们对 1 个以上节点的对应数据使用线性扩展。

RTM

逆时偏移（RTM）建模是油气勘探地震数据处理工作流程中的关键组成部分

版本

2017

加速特性

批处理算法

可扩展性

多 GPU 和多节点

SPECFEM 3D 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：288x64，100 分钟 | 为了达到与 CPU 节点相同的性能，我们对 1 个以上节点的对应数据使用线性扩展。

SPECFEM 3D

模拟地震波传播

版本

7.0.0

可扩展性

多 GPU 和多节点

更多信息

https://geodynamics.org/cig/software/specfem3d_globe

TESLA V100 性能指南

工程学

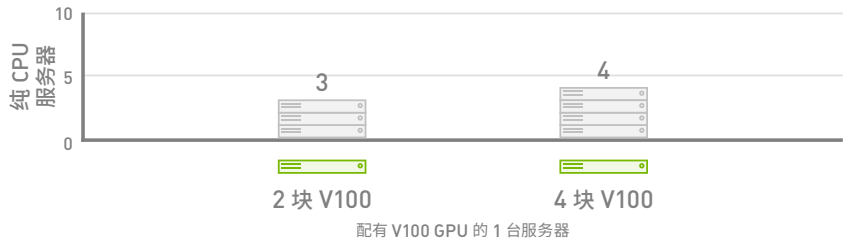
工程模拟是通过建模流程、热传递和有限元分析等开发新产品的关键环节。目前，许多常用工程应用程序均已支持 GPU 加速。数据中心配有 NVIDIA® Tesla® V100 GPU 后，在运行工程应用程序时可节省高达 20% 的服务器和基础架构购买成本，以及高达 50% 的软件许可成本。

工程适用的 TESLA 平台和 V100 的主要特性

- 对于 SIMULIA Abaqus 和 ANSYS FLUENT 等应用程序，配有 Tesla V100 的服务器可以代替多达 4 台 CPU 服务器的对应性能
- 常用工程应用程序均支持 GPU 加速
- 显存容量高达 16 GB
- 显存带宽高达 900 GB/s
- 双精度浮点运算能力高达 7.8 TFLOPS/s

SIMULIA Abaqus 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



配有 V100 GPU 的 1 台服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：7.5 | 数据集：LS-EPP-Combined-WC-Mkl (RR) | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

SIMULIA ABAQUS

用于分析结构的模拟工具

版本

2017

加速特性

稀疏直接求解器
AMS Eigen 求解器
稳态动力学求解器

可扩展性

多 GPU 和多节点

更多信息

www.nvidia.com/simulia-abaqus

ANSYS Fluent 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器

配有 V100 GPU 的 1 台服务器

CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：6.0 | 数据集：Water Jacket | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

ANSYS FLUENT

用于流体动力学模拟的通用软件

版本

18

加速特性

基于压力的耦合求解器和辐射热传递

可扩展性

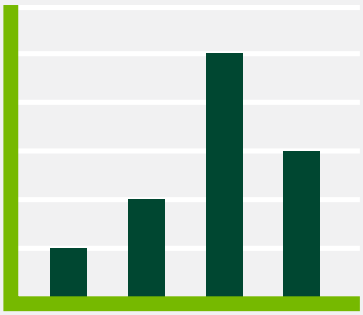
多 GPU 和多节点

更多信息

www.nvidia.com/ansys-fluent

TESLA V100 性能指南

HPC 基准测试



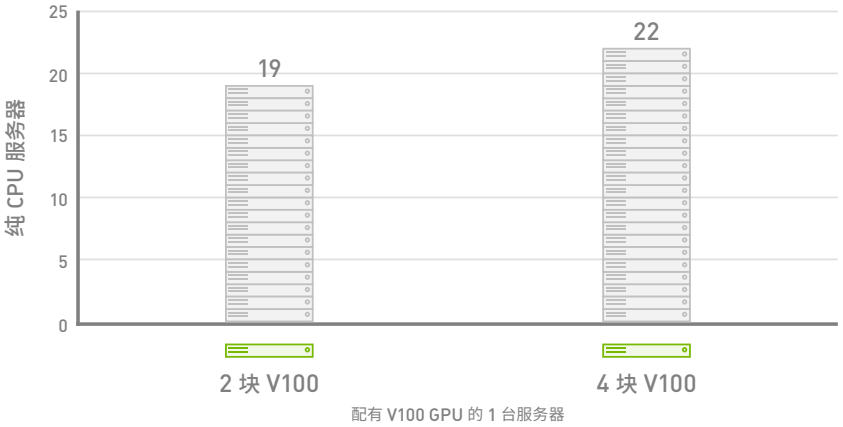
基准测试提供了系统在一定生产规模下的表现的近似情况，有助于评估不同系统的相对性能。常用基准测试具有 GPU 加速版本，可以帮助您了解在数据中心运行 GPU 的好处。

基准测试适用的 TESLA 平台和 V100 的主要特性

- 对于 Cloverleaf、MiniFE、Linpack 和 HPCG 等基准测试，配有 Tesla V100 的服务器可以代替多达 67 台 CPU 服务器的对应性能
- 常用基准测试均支持 GPU 加速
- 双精度浮点运算能力高达 7.8 TFLOPS/s，显存容量高达 16 GB
- 显存带宽高达 900 GB/s

Cloverleaf 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：bm32 | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

CLOVERLEAF

基准测试 – 迷你应用
流体动力学

版本

1.3

加速特性

Lagrangian-Eulerian
显式流体动力学迷你应用

可扩展性

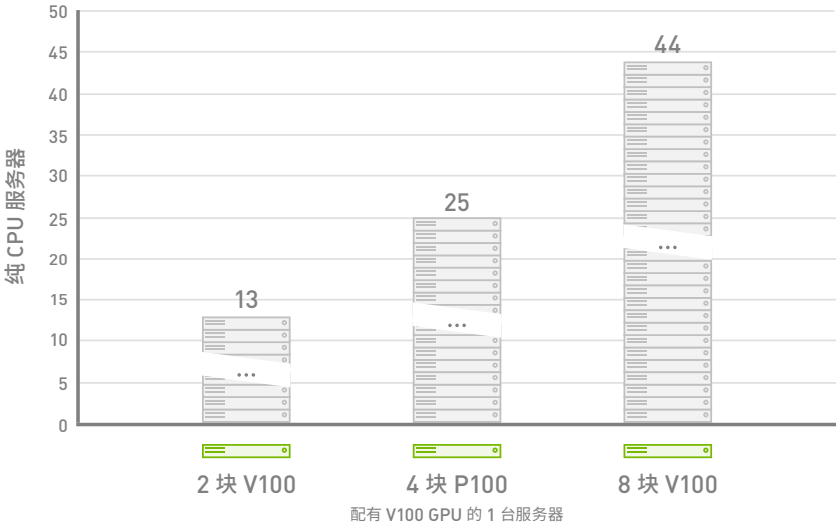
多节点 (MPI)

更多信息

http://uk-mac.github.io/CloverLeaf

MiniFE 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：350x350x350 | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

MINIFE

基准测试 – 迷你应用
有限元分析

版本

0.3

加速特性

全部

可扩展性

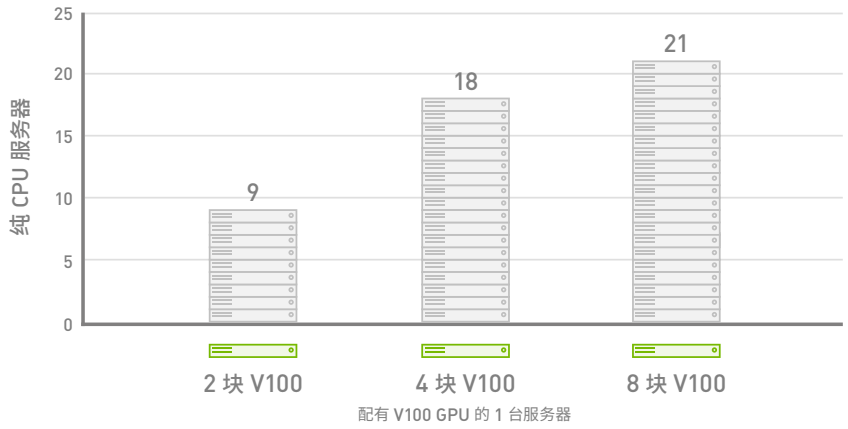
多 GPU

更多信息

https://mantevo.org/about/applications

Linpack 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 (12 GB 或 16 GB) | NVIDIA CUDA® 版本：9.0.103 | 数据集：HPL.dat | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

LINPACK

基准测试 – 测量浮点运算能力

版本

2.1

加速特性

全部

可扩展性

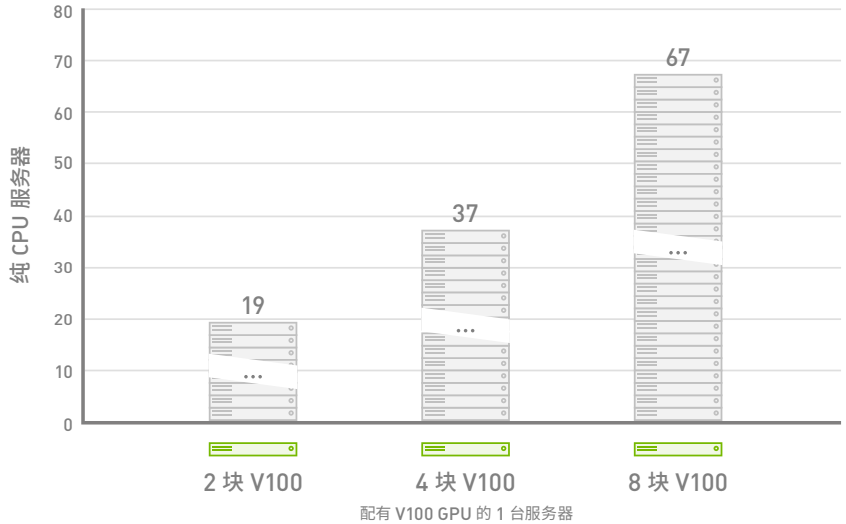
多节点和多节点

更多信息

www.top500.org/project/linpack

HPCG 性能等效图

单台 GPU 服务器与多台纯 CPU 服务器



CPU 服务器：双路至强 E5-2690 v4 @ 2.6GHz，GPU 服务器：与 CPU 服务器相同，配有 PCIe 接口的 NVIDIA® Tesla® V100 | NVIDIA CUDA® 版本：9.0.103 | 数据集：256x256x256，本地大小 | 为了达到与 CPU 节点相同的性能，我们使用了 8 个 CPU 节点时的已测量基准测试数据。然后，我们对 8 个以上节点的对应数据使用线性扩展。

HPCG

基准测试 – 练习与各种重要 HPC 应用程序密切相符的计算和数据访问模式

版本

3

加速特性

全部

可扩展性

多 GPU 和多节点

更多信息

www.hpcg-benchmark.org/index.html

TESLA V100 产品规格

	适用于 PCIe 服务器的 NVIDIA Tesla V100	适用于经 NVLink 优化服务器的 NVIDIA Tesla V100
双精度浮点运算能力	高达 7 TFLOPS	高达 7.8 TFLOPS
单精度浮点运算能力	高达 14 TFLOPS	高达 15.7 TFLOPS
深度学习	高达 112 TFLOPS	高达 125 TFLOPS
NVIDIA NVLink™ 互联带宽	-	300 GB/s
PCIe x 16 互联带宽	32 GB/s	32 GB/s
CoWoS HBM2 堆叠式显存容量	16 GB	16 GB
CoWoS HBM2 堆叠式显存带宽	900 GB/s	900 GB/s

相关假设和免责声明
常用应用程序中支持 GPU 加速的百分比数据来源于 i360 报告《HPC Support for GPU Computing》(HPC 对 GPU 计算的支持报告) 中 50 大应用程序列表。
吞吐量和成本节约的相关计算数据，为应用程序在域中以相同的计算周期进行基准测试得到的假设工作负载概要 **:http://www.intersect360.com/industry/reports.php?id=131**
匹配单个 GPU 节点所需的 CPU 节点数使用 GPU 节点应用程序加速的实验室性能结果和多 CPU 节点扩展性能进行计算。例如，分子动力学应用程序 HOOMD-Blue 的 GPU 节点应用程序加速为 37.9 倍。将 CPU 节点扩展到 8 个节点集群时，总系统输出为 7.1 倍。因此，扩展系数为 8 除以 7.1 (即 1.13)。要计算匹配单个 GPU 节点性能所需的 CPU 节点数，应将 37.9 (GPU 节点应用程序加速系数) 乘以 1.13 (CPU 节点扩展系数)，也就是需要 43 个节点。