白皮书 | 2015 年 4 月

NVIDIA® TESLA® K80 加速器

应用于科学计算与数据分析领域的全球最快加速器详解。

NVIDIA

目录

专为提供应用程序最高性能而打造

加速器在数据中心内的运用已经远超临界点。业内专家预测，今年高性能计算 (HPC) 领域中将有一半以上的新系统采用加速器，而 NVIDIA GPU 拥有 85% 的加速器市场份额[1]。人们需要更快、更准确地进行深入了解和发现，这种需求的不断增长正推动着人们越来越多地运用加速器。

Tesla K80 GPU 是用于数据分析和科学计算的全球最快加速器。Tesla K80 的设计从头到脚都是为了在现实世界应用中提供最高的性能。

就性能而言，Tesla K80 在上一代加速器的基础上实现了重大飞跃，双精度性能将近 3 TeraFLOPS (每秒浮点运算次数)，内存总带宽高达 480 GB/s。相比较与 CPU，GPU 的性能优势在 K80 身上继续增大。

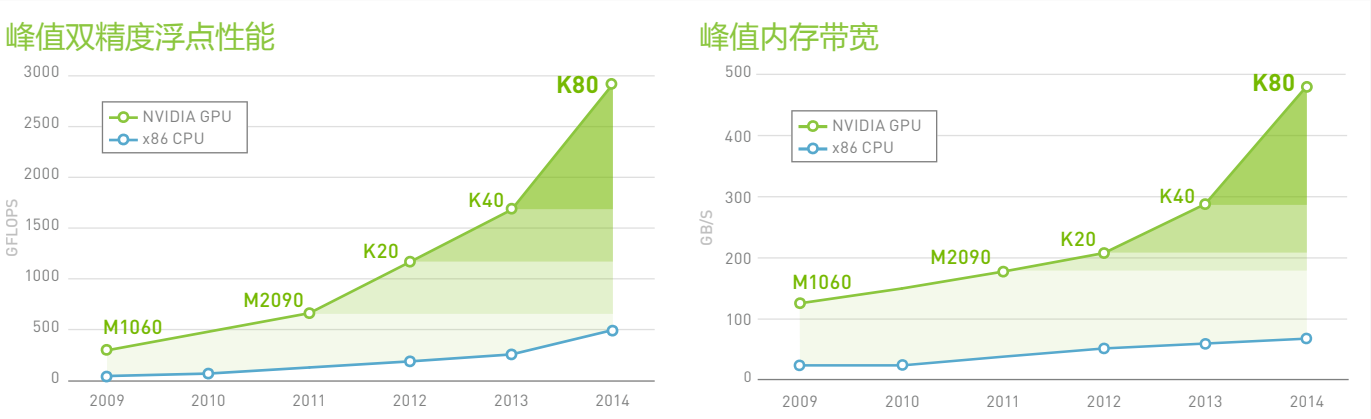


图 1: 在性能方面，Tesla K80 与上一代 Tesla 加速器和多核 CPU 相比之下实现了重大飞跃

除了浮点性能和内存带宽以外，Tesla K80 还具备旨在提供最高应用性能的诸多杰出特性。这些特性包括:

- 双核心 GPU 设计
- 改进的 GPU 动态提速技术
- 每个 SMX (流式多处理器) 配备两倍共享内存和寄存器
- 24 GB 内存总容量

[1] 2015 年高性能计算领域六大预测，2015 年 2 月特别报告，Intersect360 调研公司

专为最大限度提升应用性能而打造

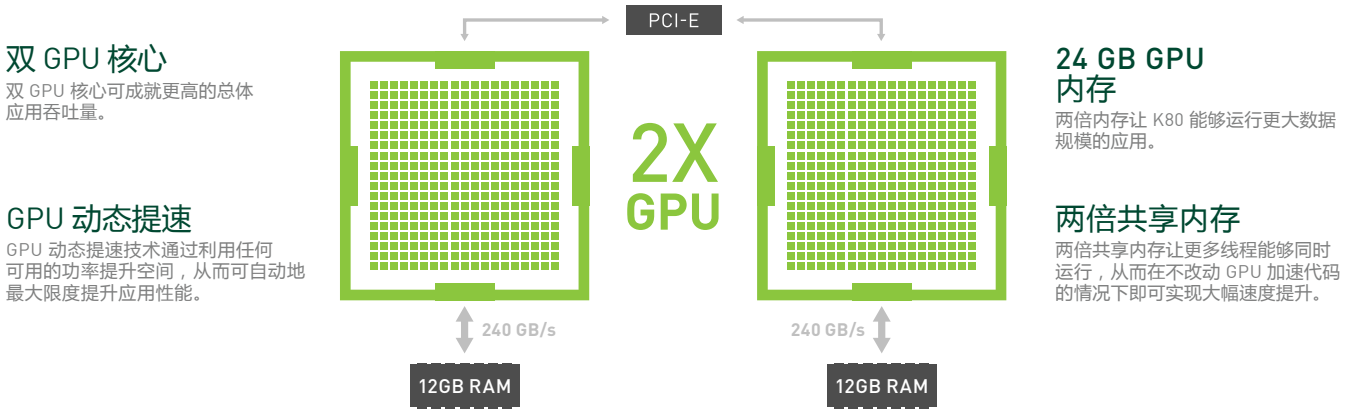
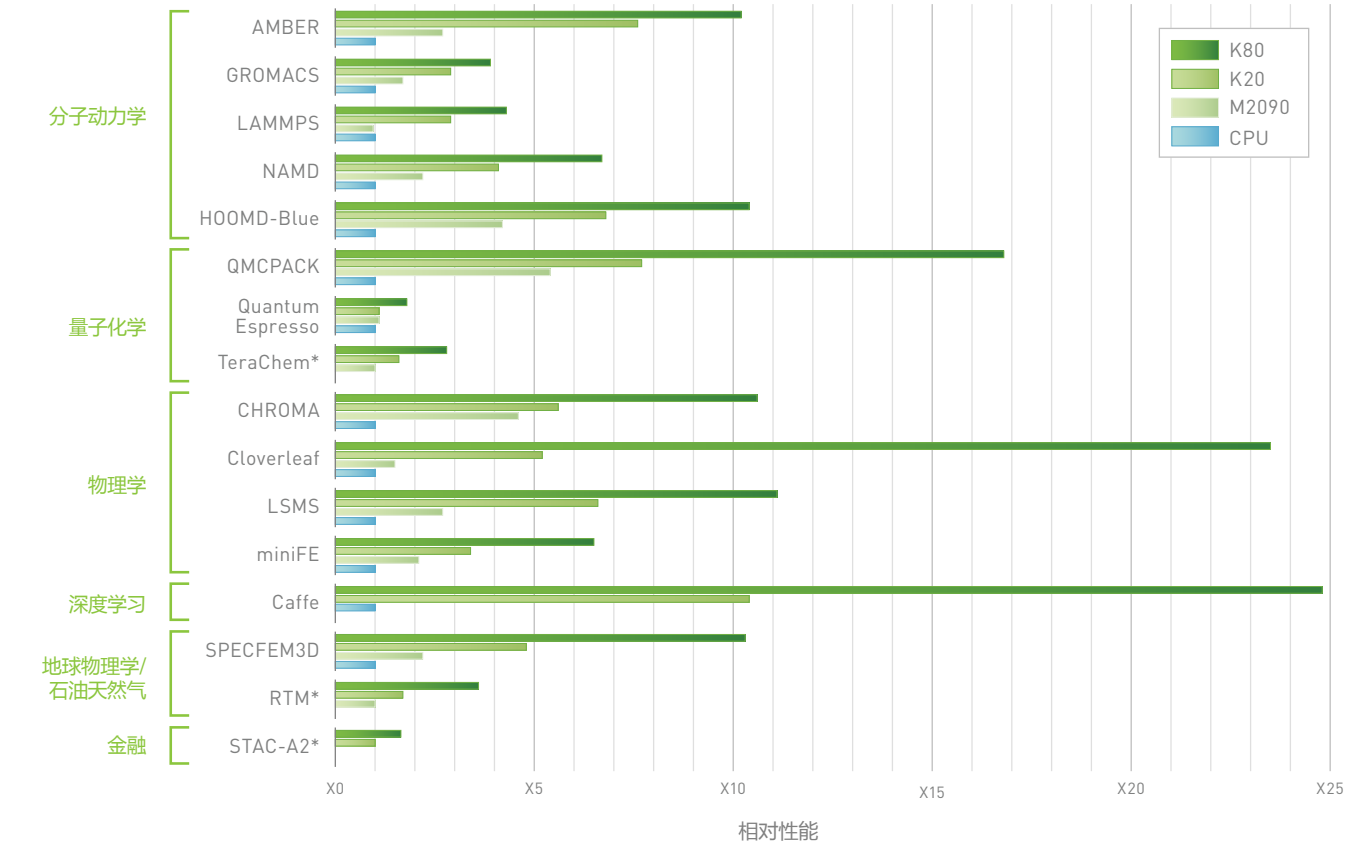


图 2: Tesla K80 的四大独特特性

最终造就了一款全新级别的 GPU 加速器，该加速器专为突破性能极限而打造。与使用 Tesla K20 时相比，高性能计算与深度学习领域中的主要应用均可实现 2 倍以上的速度提升，与基于 Fermi 架构的加速器相比可实现 4 倍速度提升。

性能比之前的产品快 2-4 倍



*没有 CPU 对比数据。| CPU 服务器: 双路 E5-2697 v2 @ 2.7GHz，GPU 服务器: 双路 E5-2697 v2 @ 2.7GHz、两块 Tesla M2090/K20/K80；K80 GPU 动态提速已启用。

图 3: K80 在各种应用上的性能比 M2090 快 4 倍、比 K20 快 2 倍。

在本白皮书中，我们将深入了解 Tesla K80，介绍它的四大特性，分析这些特性对现实世界应用性能的影响。

无论你是一名开发者还是一名运行第三方应用的研究人员，我们都鼓励你在 **www.nvidia.com/gputestdrive** 网站上注册，免费试用 Tesla K80 并对自己的应用进行基准测试。

针对数据密集型任务的双 GPU 核心

虽然浮点性能是人们广泛关注的一个性能指标，然而现实世界中的应用性能通常受限于 GPU 数据通信速度。从 Chroma 等高性能计算代码到逆时偏移 (RTM) 等能源勘探领域中的企业算法，数据从 GPU 内存到 GPU 之间的传输速度 (即内存带宽) 会直接影响应用性能的高低。

像 Tesla K80 这样的双 GPU 核心可提供一种更高效的途径来使内存总带宽，效率高于像 Tesla K40 这样的单 GPU 加速器。

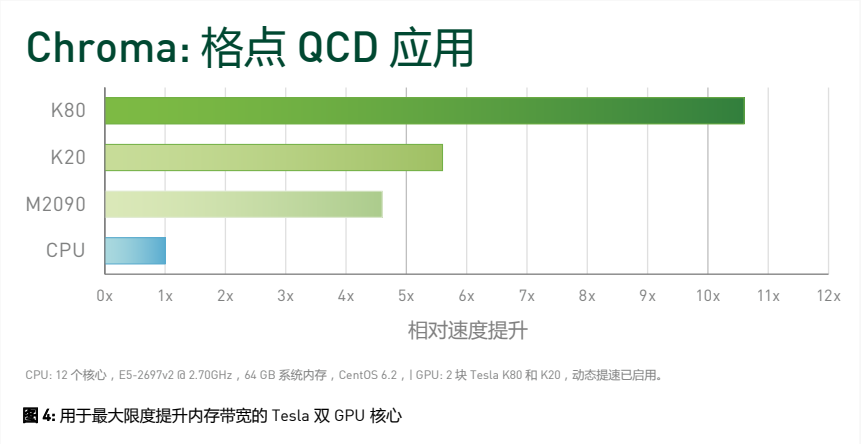


图 4: 用于最大限度提升内存带宽的 Tesla 双 GPU 核心

在 Tesla K80 中，两颗 NVIDIA® Kepler™ GPU 以 240 GB/s 的速度与 12GB GDDR5 内存相连，从而令板卡级带宽和内存容量分别达到了 480 GB/s 和 24 GB。

Chroma 是一款用来模拟粒子物理学的流行应用，其性能在很大程度上取决于内存带宽。对于像 Chroma 这样的应用，K80 可提供比其它 Tesla 加速器高 2 倍的性能。

有些客户主要关注同时运行多个任务或者关注任务吞吐量。在深度学习领域，研究人员更喜欢利用 Caffe 这样的应用来同时运行两个或更多个模型或集合。凭借双 GPU 核心，Tesla K80 可实现两倍吞吐量，从而能够大幅增强 Caffe 集合。

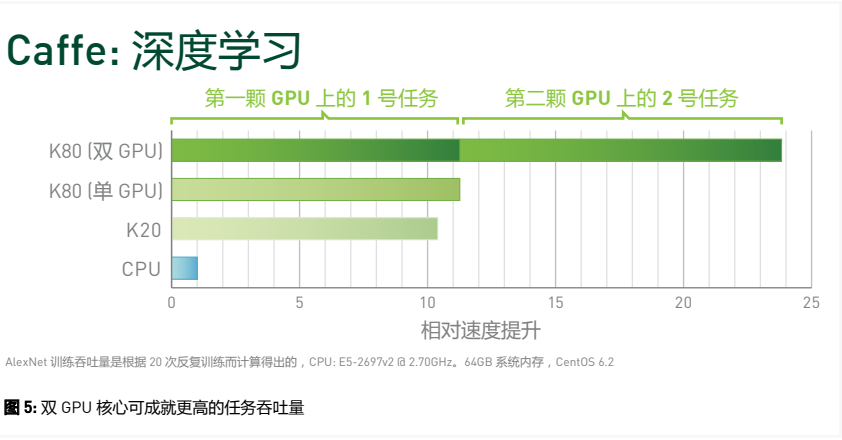


图 5: 双 GPU 核心可成就更高的任务吞吐量

GPU 动态提速: 为每一款应用提供最高性能

Tesla K80 中的 GPU 动态提速技术现已重新设计，以便无缝且智能地为任何特定应用提供最快的性能。通过将核心时钟频率提升至最高水平，同时不超出 GPU 固定的功率预算，现实世界中的应用能够实现 40% 以上的速度提升，GPU 的利用率也大大提升。

无论是 CPU 还是 GPU，每一种处理器都是针对特定功率预算而设计的，这种预算即热设计功率 (TDP)。额定的热设计功率是功率上限，K80 的热设计功率为 300W。虽然像 Linpack 这样的少数几款浮点运算密集型应用在最低时钟频率设置或基础时钟频率下会达到峰值热设计功率，但是大多数高性能计算任务在这些设置下不会接近功率预算值。对这些应用来说，K80 能够智能地设置最佳的时钟频率，同时不超出 300W 的极限，因而能够为任务提供最快的计算速度。

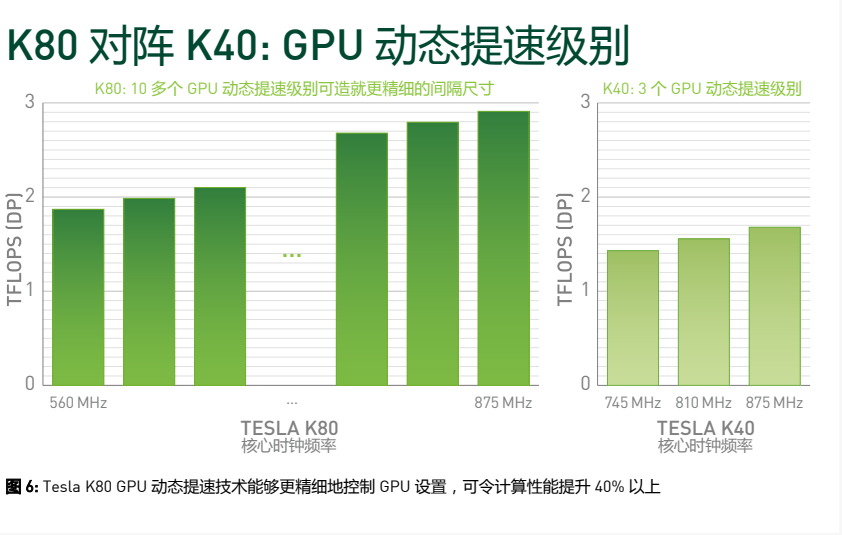


图 6: Tesla K80 GPU 动态提速技术能够更精细地控制 GPU 设置，可令计算性能提升 40% 以上

虽然 GPU 动态提速技术首次出现是在 Tesla K40 上，但在 K80 中重新设计的 GPU 动态提速技术支持更大范围的时钟频率，在基础频率和最高动态提速频率之间有 10 多个级别。在对比动态提速打开与关闭之间的性能差异时，利用 GPU 动态提速技术，使用最高主频设置，GPU 浮点计算性能可提升 40% 以上，同样对现实应用程序性能也能提升 40% 以上。

在 Tesla K80 中，GPU 动态提速特性默认情况下是打开的。如需关闭动态提速，只要使用 nvidia-smi 实用程序即可。

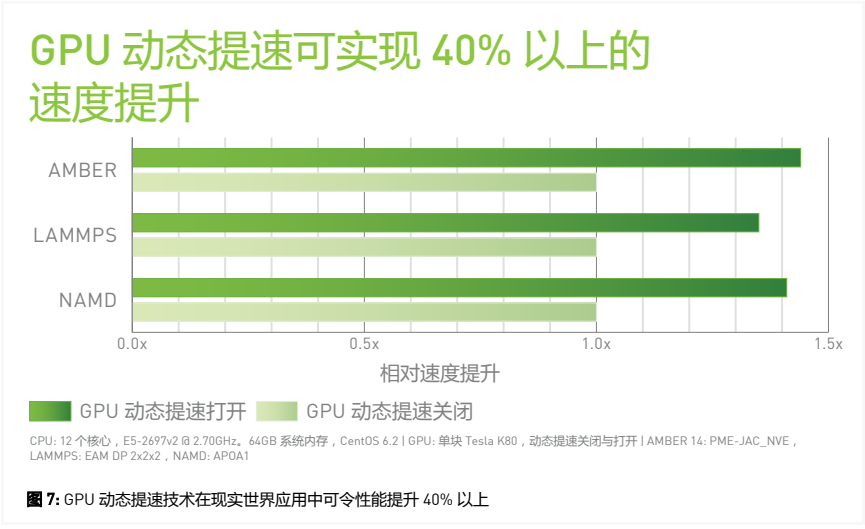


图 7: GPU 动态提速技术在现实世界应用中可令性能提升 40% 以上

双倍共享内存和寄存器可提升计算效率

Tesla K80 GPU 架构被称作 GK210，它继承了 Kepler 架构的所有优点，例如节能的 SMX、动态并行机制以及 Hyper-Q 技术。然而 GK210 独有的一大特性是每个 SMX 的共享内存和寄存器文件容量更大。

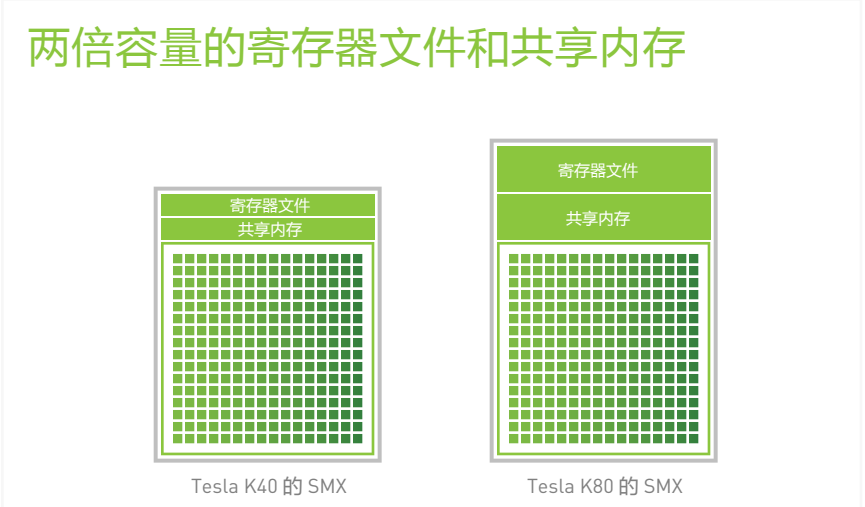


图 8: 与较旧的 Kepler GPU 相比，Tesla K80 拥有两倍容量的寄存器文件和共享内存

通过降低数据溢出或数据抖动的风险，使并发量翻倍有助于掩盖算术延迟和内存延迟，从而提升整体效率与应用性能。更多的寄存器和共享内存让开发者能够自动加速他们的应用，加速幅度最高可达 30%，开发者只需启动更多的并发线程即可实现加速。无需改动代码。编译器能够智能地识别 Tesla K80 何时有空启动更多的并发线程。

例如，逆时偏移 (RTM) 是一种用于地震处理的数据密集型算法，该算法受内存带宽和寄存器占用率的制约。与 Tesla K40 相比，Tesla K80 可提供 97% 的应用整体速度提升。在这 97% 当中，30% 的速度提升归功于每个 SMX 更大容量的寄存器文件。

请注意，GK210 具备的计算能力 (Compute Capability) 为 3.7。

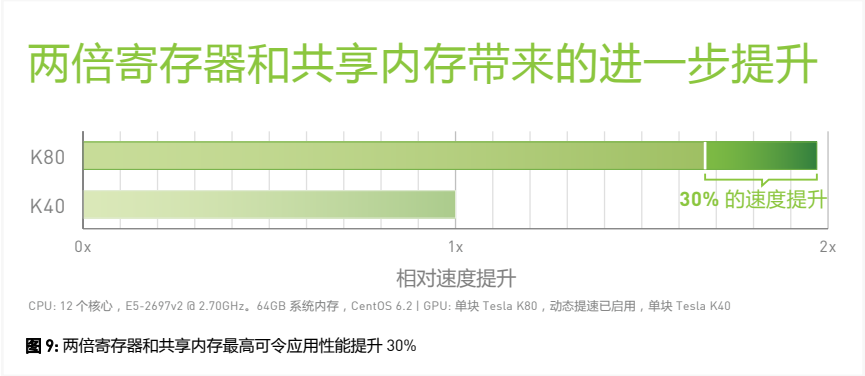


图 9: 两倍寄存器和共享内存最高可令应用性能提升 30%

总容量 24 GB 的内存可满足数据密集型的任务需求

许多高性能计算和数据分析任务需要大型模型加载到 GPU 的内存中。如果数据集无法载入到可用的 GPU 板载内存之中，那么就需要和系统内存直接进行频繁的数据交换，应用性能会大打折扣。

Tesla K80 是首款提供内存总容量为 24 GB 的加速器，这一容量是 Tesla K40 的两倍，比 K40 以前的任何其它 GPU 大四倍。在软件中，代码必须能够运行在多个 GPU 核心上，以便把工作任务分布到两个 12 GB 内存分区当中。

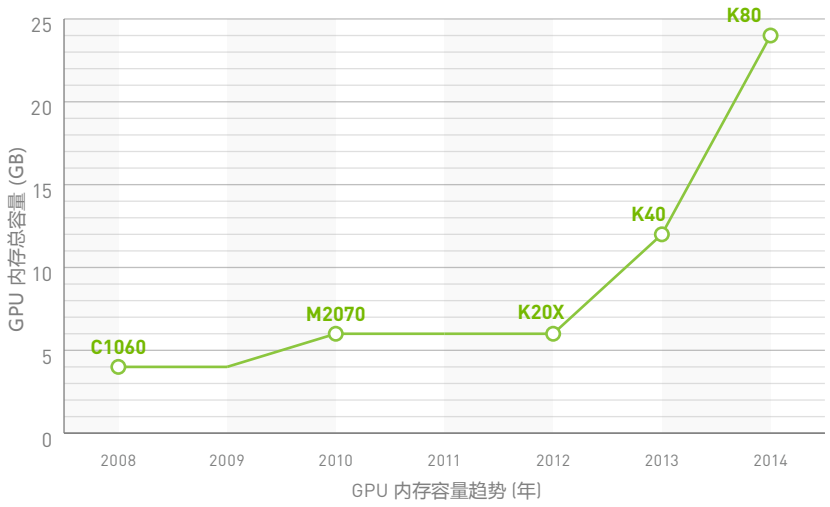


图 10: Tesla K80 的内存容量是 K40 的两倍，比其它 GPU 大四倍

石油天然气、制造业以及深度学习等应用都是需要大容量 GPU 内存的例子。ANSYS ® Fluent ® 是制造业中最流行的商业应用，该应用能够大大得益于 Tesla K80。一个拥有 1400 万个单元的卡车模型需要使用 65 GB 的 GPU 内存，或者至少需要使用四个服务器节点，每个节点配备两块 Tesla K40 加速器。如果使用 Tesla K80 的话，同样的任务只需两个服务器节点即可，每个节点配备两块 Tesla K80 加速器。

由于把服务器数量缩减了一半，Tesla K80 的性能略高于 Tesla K40 服务器，因而让客户能够体验到生产率的大幅提升。凭借四台 Tesla K80 服务器，客户每天运行的工作量可达四台 Tesla K40 服务器的两倍。

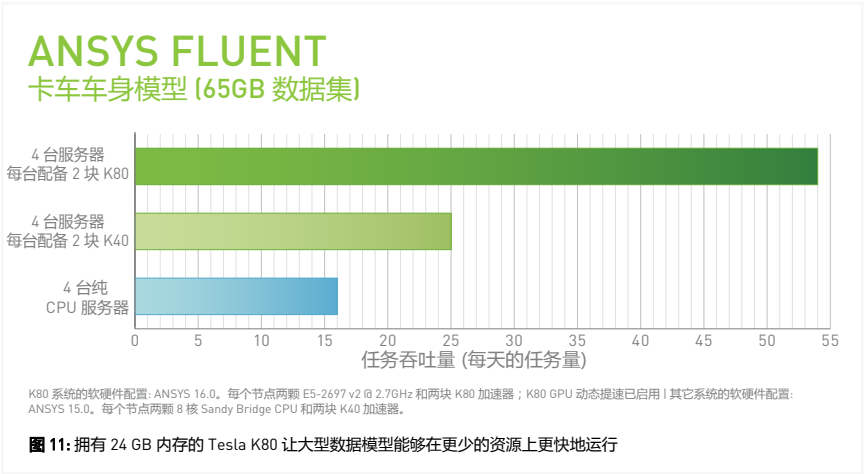


图 11: 拥有 24 GB 内存的 Tesla K80 让大型数据模型能够在更少的资源上更快地运行

结语

NVIDIA Tesla K80 的设计从头到脚都是为了给现实世界应用提供最高的计算性能。除了 Kepler 架构诸多的先进特性以外，K80 还拥有最大限度专注于提升应用程序性能的四大全新特征：

- 双 GPU 核心
- 改进的 GPU 动态提速技术
- 每个 SMX (流式多处理器) 配备两倍共享内存和寄存器
- 24 GB GPU 内存内存总容量

从受带宽制约的应用到受计算性能制约的应用，K80 可在所有使用场合提升客户的生产率。

我们鼓励大家参加 K80 试用计划，免费试用 Tesla K80 对自己的 GPU 应用程序进行基准测试，体验应用程序在 K80 平台带来的性能大幅提升，申请网址为：**www.nvidia.com/gputestdrive**，鼓励大家在自己的应用中体验大幅速度提升。

如需了解 Tesla K80 以及 Kepler 加速的更多通用细节，我们鼓励大家访问 **http://docs.NVIDIA.com/cuda/kepler-tuning-guide/index.html**