白皮书

使用 NVIDIA 虚拟 GPU 支持混合工作负载

提高组织工作效率

2019 年 4 月

NVIDIA

目录

概览 3

什么是混合工作负载? 5

混合工作负载详解 6

部署混合工作负载 12

结语 12

概览

当今专业应用程序的要求很高，这意味着企业中的 GPU 用例要比以往更多。设计师和工程师依赖具有 3D 可视化功能的图形密集型应用程序，其中许多内含 AI 增强功能。数据科学家运行由 AI、深度学习和推理提供支持的计算密集型应用程序。即使是知识型工作者，现在也越来越多地采用图形密集型办公效率应用程序 – 最近仅 Windows 10 更新已使 GPU 的使用增加了 20%。使用 GPU 可快速处理数据，让这些应用程序的速度呈指数级提升，从而提高性能并帮助公司更快地进行创新。

但是，使用高度专业化应用程序的专业人士仍然会在工作效率方面遇到障碍。由于数据中心效率低下，物理工作站存在诸多限制，许多工作流程往往会变慢。每当用户下载或上传大型文件、提交文件以供渲染或等待提交到高性能计算 (HPC) 集群的作业的结果时，他们都会白白浪费宝贵的时间。他们要么等待完成计算能力远超其端点能力的进程，要么等待在数据中心内进行的、可从访问未开发资源中受益的工作执行完毕。最终都会导致上市时间延迟。

企业现在正在寻找能够显著加快这些专业工作流程的 IT 解决方案，以便在会议结束或工作日结束之前做出决策。IT 专业人士需要部署灵活的 IT 基础设施（无缝提供更高的 GPU 利用率）来为企业提供支持，从而以更少的资源和投资满足更多业务需求。

利用 NVIDIA 虚拟 GPU 和混合工作负载加速工作流程。

实现这些目标的极佳方式是利用虚拟化，将 NVIDIA 虚拟 GPU (vGPU) 软件与业内功能非常强大的 NVIDIA® GPU 结合使用。因为相同的 NVIDIA GPU 可用于虚拟机 (VM) 和深度学习、AI、推理、高性能计算和其他应用程序，所以 IT 可以将多种工作负载融合在一起，以提高数据中心效率。由此产生的混合工作负载可确保数据中心的 GPU 资源得到充分利用，同时加快专业工作流程的速度。

使用通用服务器资源运行高性能计算和虚拟化工作负载可使灵活性和运营效率更上一层楼。借助混合工作负载，工程师和设计师可以在白天将计算资源用于虚拟桌面架构，然后在晚上重复这些资源来运行计算作业，从而大大提高利用率。您无需再管理物理工作站，因此 IT 成本也会随之降低。

虚拟化工作站可以确保用户在不影响性能的情况下获享移动性和安全性。例如，VM 用户实际上并不会像在物理工作站中工作时那样下载模型或数据。VM 能在数秒之内访问数据中心内存储的大型文件和数据集，从而大大提升工作效率。因为数据并未实际下载或上传，所以项目团队还可以避免版本控制方面的问题。

从用于产品设计的逼真渲染到用于制造业的数据密集型模拟，再到速度实现指数级提升的媒体和娱乐内容渲染，混合工作负载都加速了专业工作流程，从而提升运营效率，包括提高输出精度、缩短上市时间以及提高用户工作效率。

什么是混合工作负载?

并非所有工作负载都是相同的。如今，大多数应用程序都属于范围广泛的可视化和计算领域。这两个领域包含多个子领域，而每个子领域都有各自独特的需求。

- **办公效率应用程序**对 GPU 加速的需求与日俱增，原因在于应用程序的要求日益严苛，而且企业都逐步迁移到 Windows 10。
- **2D 电子设计自动化 (EDA) 应用程序**通常需要 GPU 加速才能在虚拟桌面架构环境中出色运行。
- **3D 专业应用程序**需要 GPU 加速（例如，Dassault Systèmes CATIA 和 SOLIDWORKS）。
- **深度学习、AI 和推理**需要功能强大的 GPU 才能实现更高的效率。

通过在数据中心内配备 NVIDIA GPU，企业现在可以运行多种类型的工作负载。无论是用于评估金融投资的蒙特卡罗模拟，还是用于油气勘探的 3D 图形和数据密集型工作负载，均可通过重复利用主机来使用相同的基础设施，在白天运行虚拟桌面架构，而在夜间运行高性能计算和其他计算工作负载。

迁移 GPU 加速的 VM

得益于实时迁移，混合工作负载成为可能。实时迁移是指将运行中的 VM 从一个物理主机系统迁移到另一个物理主机系统，而最终用户无需中断操作，也不会造成数据损失。实时迁移技术已问世多年，但直至近期，我们才实现 GPU 的实时迁移。NVIDIA 虚拟 GPU 软件是业内率先支持（也是迄今为止唯一支持）实时迁移 GPU 加速的 VM 的技术。

迁移包含 GPU 加速技术的 VM 是一项艰难的任务。CPU 仅包含几个核心，而 GPU 却包含数千个核心。实时迁移必须将 GPU 从一台服务器复制到另一台服务器上，将其进程一一映射，同时还必须复制正在使用的所有活跃组件的状态。

利用 GPU 实时迁移实现混合工作负载不是天方夜谭，这是因为 NVIDIA 的虚拟化软件（NVIDIA Quadro® 虚拟数据中心工作站 (Quadro vDWS) 和 NVIDIA GRID®）可以作为深度学习、推理、训练和高性能计算工作负载在相同的 Turing™、Volta™ 和 Pascal™ GPU 上运行。现在，借助 NVIDIA vGPU 软件和 VMware vMotion 或 Citrix XenMotion，IT 可以提高数据中心的敏捷性，并在数秒之内实时迁移用户。因此，IT 能更大限度地提高数据中心利用率，同时还能更高效地维护服务器。

更大限度地提高数据中心利用率。 IT 管理员可以随时将实时 VM 整合到未充分利用的服务器节点。当每天工作结束，用户陆续回家时，IT 可以将剩余 VM 实时迁移到其他主机，从而实现整合。然后，他们可以在夜间重复利用原始主机运行计算工作负载，例如高性能计算和深度学习。次日，当虚拟桌面架构需要图形资源时，IT 管理员可以再一次轻松地将 NVIDIA GPU 重复利用到虚拟 GPU，以此为虚拟桌面架构提供支持。

让服务器保持正常运行。 通过迁移 VM，IT 可以独自执行诸如工作负载平衡、基础设施恢复和服务器软件升级等关键服务，而不会造成任何 VM 停机。这样可确保服务器始终保持出色的工作状态，最终用户永远不会遭遇任何中断或数据损失。

其他优势：

- **业务中断减到最少。** 无需安排停机即可执行服务器维护（如硬件更换 / 升级或软件更新）。
- **改善服务器密度。** 手动地在用户之间平衡负载，并整合常用的配置文件类型，从而改善密度。
- **优化基础设施使用情况。** 将 VM 迁移到另一台主机，以允许在下班后使用主机和 GPU 进行计算。
- **提高敏捷性。** 利用在迁移期间从 NVIDIA 虚拟 GPU 软件监控中获取的重要信息，确保在高可用性的情况下提供优质的用户体验。

混合工作负载详解

通过深入研究典型的计算机辅助工程 (CAE) 工作流程，可以更好地理解实施混合工作负载的优势。在本示例中，某位工程师正在建造一辆赛车，同时在该项目中，该工程师将确定赛车的结构能否在安全情况下应对实际场景的作用力，从而验证采用尽可能轻的材料进行设计是否可行。以下是他的工作流程：

1. **首先进行 CAD 设计。** 赛车的完整组件将在 CAD 应用程序中显示。在本示例中，要分析和验证的结构是赛车的主轴几何结构。

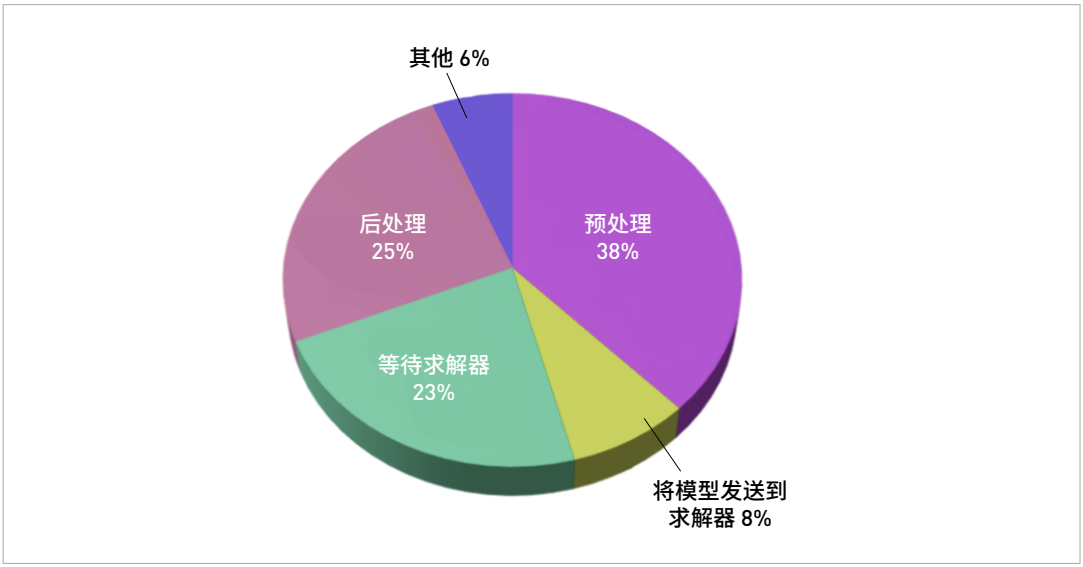
2. **预处理。** 此流程将测试导入的几何结构是否存在错误，并对其进行简化以供模拟（以确定是否存在任何锐边等）。下一步是构建一个有限元网格，以离散几何结构。此流程称为网格化。即使最小的模型部分（例如主轴）也可能包含多达两百万个元素，这些元素反过来会转化为多个自由度。此时，工程师将为几何结构分配材料并应用限制和作用力。

3. **高性能计算求解器。** 接下来，高性能计算求解器会对所构建的有限元模型求解。这是一项需要大量计算的作业，一次模拟可能需要数小时甚至数天时间。在此方面，NVIDIA GPU 加速可以帮助提高计算效率并缩短作业周转时间，以更快的速度提供结果，从而在相同的投资条件下提高许可证的利用率。

4. **后处理。** 完成分析后，工程师将对结果进行分析和推理以验证设计，检查结构的完整性并对设计 / 有限单元模型进行必要的更改。然后，重新运行模拟。

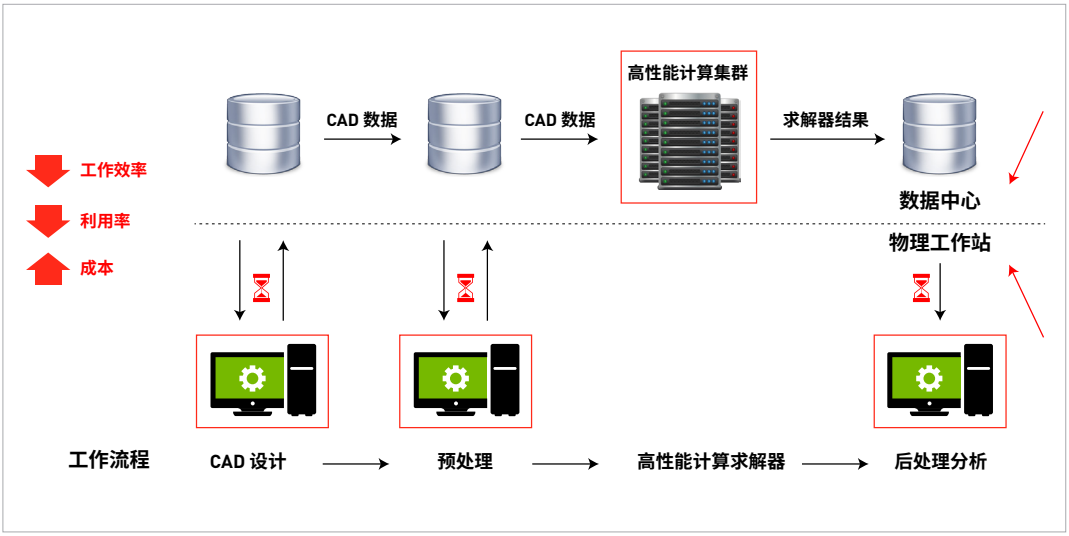
当工程师执行此工作流程时，他们会使用多种应用程序。他们会先保存自己的文件，然后再离开某一应用程序以启动另一工作流程。之后，他们将返回检查，并来回切换到不同的应用程序，每个应用程序都需要独特数量的计算和可视化资源。为更大限度地提高效率，工程师通常会设计并提交模型，以便在工作日期间进行多次预处理。晚上，他们会将模型提交到高性能计算求解器，以便在第二天早上他们到达办公室时，结果准备就绪，可以进行分析。

典型工作流程的每个阶段要花费多少时间?多项研究表明，工程师需要在设计、预处理和后处理阶段花费大约 66% 的时间。基本上而言，这意味着这几个阶段极具交互性，此时工程师需要坐在自己的工作站旁，积极处理模型。在剩余的三分之一时间，高性能计算求解器会执行分析，这种分析不具备交互性。



传统部署会浪费宝贵的资源。

在传统部署中，工程师通常会在 CAD 工作中使用物理工作站。数据中心通常存在高性能计算集群和存储空间，用于协作目的。当工程师想处理 CAD 数据文件时，他们会将其签出。然后，他们可以在准备好进行预处理时将文件签回。签回之后，数据将被传递到高性能计算集群，由高性能计算求解器分析结果。随后，这些结果将导入回工程师的工作站，以执行进一步分析。

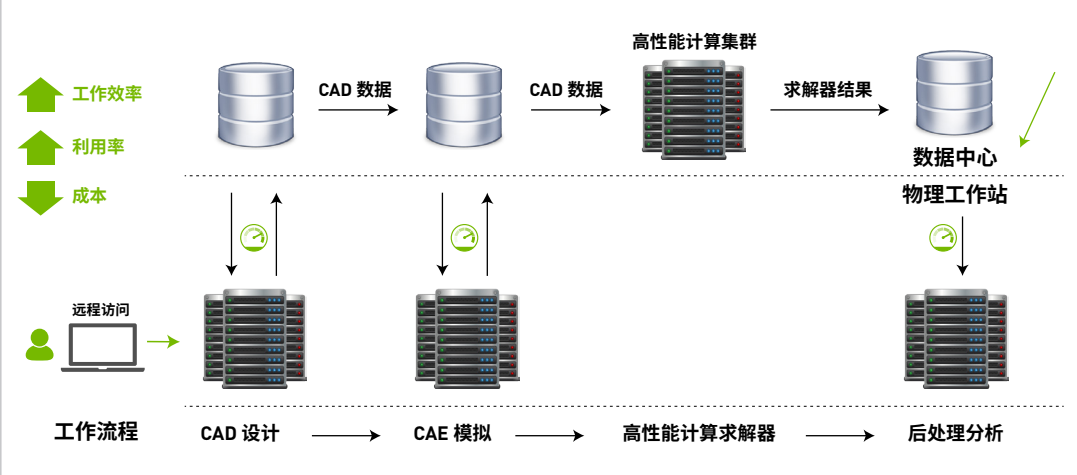


此类部署存在不少缺点和限制。首先，工作效率会受到负面影响。根据所执行的分析类型，高性能计算求解器通常会生成包含数 TB 数据的文件。其他阶段会生成数十到数百 MB 的数据。因此，在文件上传和下载方面，工程师可能要花很长时间。根据工作站的位置和文件大小，文件访问可能需要 1 到 20 几分钟不等。因为这些模型在一天内要经历大量迭代，所以工程师需要反复上传和下载文件。从对大型工程团队的工作效率的累积影响方面考虑，不难看出大量时间都花费在了等待上。

工作效率仅仅是浪费的一方面，利用率也存在此类问题。在工作日期间，工作站的利用率很高，而高性能计算集群可能未得到充分利用。当工程师每天结束工作，下班回家后，他们的工作站处于空闲状态，而高性能计算集群处于高强度利用状态。此外，成本也很高。IT 部门必须同时维护和管理数据中心以及物理工作站。也就是说，IT 人员将所有时间都用来采购、部署、升级和修补两套硬件和软件。

虚拟化部署可以消除限制。

不妨考虑一下，如果将工作站迁移到数据中心并虚拟化此环境，会对工作效率、利用率和成本产生哪些影响。



工作效率可立即提升。由于工作站现在靠近托管 CAD 文件的存储位置，因此工程师无需等待文件下载，而是在数秒之内即可访问。此外，存储访问的吞吐量也极高，因此即使工程师反复迭代文件，他们仍能以极短的等待时间高效地工作。

另一项优势是利用率提高。在工作日期间，当高性能计算集群处于空闲状态时，计算资源可以由正在积极处理 CAD 设计的工程师使用。从本质上讲，高性能计算集群可重复用于预处理和后处理。最后，IT 部门不用再管理物理工作站，而只需管理数据中心即可。 通过提高利用率和减少要管理的资源，减少了支出并降低了成本。

现在，我们来看看这种混合工作负载概念在数据中心内是如何发挥作用的。

GPU 加速的虚拟化的优势

1. 可节省上传和下载大型模型和数据文件的时间

   曾经需要 8-16 小时处理时间的项目，现在的处理时间可缩短到大约 40 分钟。

2. 运行混合工作负载以实现持续的资源利用

3. 降低资本支出 (CAPEX) 和运营支出 (OPEX)

   物理工作站的成本可能超过 10,000 美元。这一成本是虚拟工作站成本的两倍多。

4. 在数据中心中安全地保存数据，以便随时随地通过任何设备远程访问数据

混合工作负载如何充分利用资源。

以下是在数据中心内实施混合工作负载的方式。在下图中，两个服务器节点各自安装了两个运行 NVIDIA 虚拟 GPU 软件的 Tesla 数据中心 GPU。这些服务器用于托管虚拟桌面架构环境，并在工作日期间运行多个 VM。

VM 正在运行 Windows 10 和 Office 办公效率应用程序、2D EDA 应用程序，以及 3D CAD 和 CAE 工作负载。在一整天内，所有这些工作负载均由相同的计算节点托管。如果用户在当天结束时开始注销自己的 VM，两个节点开始变成未充分利用状态。过去，如果计算资源未被充分利用，则无法使用这些备用资源。如果您希望运行高性能计算工作负载并获得快速结果，务必要利用这些额外资源。

现在，借助 NVIDIA vGPU 实时迁移，当用户下班注销时，您可以实时迁移任何剩余的 VM，将其整合到一个节点上。

在此过程中，您将释放第二个节点，并可立即重复利用该节点，以便在夜间运行高性能计算求解器工作负载。

第二天早上，当高性能计算求解器完成工作且用户回来上班后，VM 可以重新上线，并且相同的节点可用于后处理。

部署混合工作负载

充分提高数据中心的效率和敏捷性

现在，您已经了解在数据中心内运行混合工作负载的优势。要开始使用，您需要满足以下要求：

- **NVIDIA 数据中心 GPU。** 在业界，NVIDIA 数据中心 GPU 具备超凡的性能，可在计算工作负载和虚拟化工作负载中使用。
- **NVIDIA 虚拟 GPU 软件。** 借助 NVIDIA vGPU 软件，每个 VM 均能使用 NVIDIA GPU 的强大功能，从而确保提供更出色的性能和非凡的用户体验。此外，它还提供了适当调整 VM 大小的功能。
- **持续支持正常运行。** 通过由 GPU、NVIDIA vGPU 技术以及 VMware vMotion 或 Citrix XenMotion 支持的实时迁移，确保您充分利用环境并使环境的正常运行时间达到极高水平。
- **端到端的监控和见解。** 您的工作负载非常复杂。NVIDIA vGPU 软件可提供与多个管理和监控工具（例如 VMware vROPS 和 Citrix Director）的集成，从而在整个虚拟化堆栈中提供端到端的见解。分别在主机、VM 和应用程序级别获取对 GPU 的管理和监控，以便快速解决问题及进行前瞻性管理。

结语

混合工作负载之所以能够实现，是因为运行 NVIDIA vGPU 软件（NVIDIA Quadro vDWS 和 NVIDIA GRID）与运行深度学习、推理、训练和高性能计算工作负载的 Turing、Volta 和 Pascal GPU 相同。通过在数据中心内配备 NVIDIA GPU，IT 可以使用实时迁移随时将 VM 整合到未被充分利用的服务器节点，以重复利用主机，并确保高性能计算和其他计算工作负载始终能够访问尽可能多的资源。通过迁移 VM，IT 可以执行工作负载均衡、基础设施恢复和服务器软件升级等关键服务，而不会造成任何 VM 停机或数据丢失。借助混合工作负载，IT 可以充分利用数据中心资源，同时提供高可用性和优质用户体验。

要了解如何在您的环境中部署 NVIDIA 混合工作负载，请参阅我们的参考设计指南：混合工作负载 – 常见架构上的虚拟桌面和高性能计算。

详细了解 NVIDIA 如何提高数据中心的效率。