NVIDIA 虚拟计算服务器
利用虚拟 GPU 助力计算密集型工作负载

变革虚拟计算

随着数据中心服务器数量的增长，IT 管理员期望使用来自 VMware、红帽、Nutanix 和 Citrix 的标准服务器虚拟化平台管理服务器。如 Gartner 所述，“基于虚拟化管理程序的服务器虚拟化技术现已发展成熟。对于大多数中到大型企业而言，其 80% 到 90% 的服务器工作负载都在虚拟机 (VM) 中运行。1” 然而，除虚拟桌面基础架构外，使用基于虚拟化管理程序的虚拟化技术的这一传统数据中心架构皆局限于 CPU 服务器。基于这一原因，运行 AI、深度学习、数据科学和高性能计算 (HPC) 工作负载的 GPU 加速服务器通常位于数据中心的其他服务器中，与 CPU 服务器隔离，进而限制了利用率、灵活性和可管理性。

NVIDIA® vComputeServer 使 GPU 加速服务器能够拥有基于虚拟化管理程序的服务器虚拟化技术的优势。数据中心管理员现在可以在虚拟机 (VM) 中使用 GPU 来运行各种计算密集型工作负载。

vComputeServer 软件可通过虚拟化 NVIDIA GPU 来加速大型工作负载，软件中包含 600 多个用于 AI、深度学习和高性能计算的 GPU 加速应用。GPU 共享可以让单个 GPU 助力多个 VM，从而更大限度地提高利用率和可购性，也可实现多个虚拟 GPU 助力单个 VM，从而完成十分密集的工作负载。凭借对所有主要的虚拟化管理程序虚拟化平台的支持，数据中心管理员可以使用与其他数据中心相同的管理工具来管理 GPU 加速服务器。

计算许可

与 NVIDIA® GRID® vPC/vApps 和 Quadro® 虚拟数据中心工作站 (Quadro vDWS) 不同，vComputeServer 不与单个并发用户绑定，而是为每个物理 GPU 提供 1 年期的订阅许可，其中还包含 NVIDIA 企业支持。如此，多个 VM 中的大量计算工作负载便可在单个物理 GPU 上运行，进而更大限度地提高利用率和投资回报率。

使用 NGC 软件优化容器

vComputeServer 支持用于深度学习、机器学习和高性能计算的 NVIDIA NGC GPU 优化软件。NGC 软件包括为顶级 AI 和数据科学软件打造的容器，由 NVIDIA 进行调整、测试和优化。此外，该软件还包括用于高性能计算应用和数据分析的经过全面测试的容器。

NGC 还为各种常见的 AI 任务提供预训练模型，这些模型已经过针对 NVIDIA Tensor Core GPU 的优化，并包括用于创建具备样本性能和准确性指标的深度学习模型的指令和脚本。这可以帮助数据科学家、开发者和研究人员减少部署时间和降低项目复杂度，使其专注于构建解决方案、收集见解以及提供业务价值。

特性

- **GPU 性能** - 在虚拟环境中访问功能强大的 GPU。
- **管理和监控** - 利用基于虚拟化管理程序的工具简化数据中心的可管理性。
- **实时迁移** - 无需中断，便可实时迁移 GPU 加速 VM，简化维护和升级过程。
- **更大限度地提高利用率** - 通过 GPU 共享和多个 GPU 聚合来提高利用率和工作效率。
- **安全** - 将服务器虚拟化的优势扩展到 GPU 工作负载。
- **多租户** - 分离工作负载，从而为多个用户提供安全支持。
- **快速部署** - 利用经 GPU 优化的 NGC 容器完成 AI、数据科学和高性能计算任务。
- **可靠性** - 使用纠错码 (ECC) 和动态页面引退防止数据发生损坏。
- **企业软件支持** - 企业软件支持 - 通过 NVIDIA 企业和 NVIDIA NGC 支持服务获得支持。

NVIDIA vCOMPUTESERVER 功能列表

配置和部署	
GPU 共享（部分）	√
GPU 聚合（多个 vGPU）	√
通过 NVLink 提供对等通讯	√
纠错码和动态页面引退	√
支持 Linux 操作系统	√
支持 Windows 操作系统	x
NVIDIA 计算驱动程序	√
NVIDIA 图形驱动程序	x
NVIDIA Quadro 驱动程序	x
服务质量调度	√

数据中心管理	
主机、客户机和应用级监控	√
实时迁移	√
支持	
NVIDIA 直接提供的企业级技术支持	√
维护版本、缺陷修复程序和安全补丁 [2]	√
NGC 支持服务 [3]	√

vCOMPUTESERVER 配置文件

支持的最大帧缓存容量	48GB
支持的最小帧缓存容量	4GB
最大多租户比例	8:1
可用配置文件	4C、6C、8C、12C、16C、24C[4]、32C[5]、48C[6]

推荐用于 vCOMPUTESERVER 的 GPU

	NVIDIA T4	NVIDIA V100 (SXM2)
RT 核心数	48	-
Tensor Core	320	640
CUDA® 核心数	2,560	5,120
内存	16 GB GDDR6	32 GB HBM2
FP 16/FP 32（混合精度）	64 TFLOPS	125 TFLOPS
FP 32（单精度）	8.1 TFLOPS	15.7 TFLOPS
FP 64（双精度）	-	7.8 TFLOPS
NVLink：每个 VM 的 GPU 数量	-	最多 8 个
纠错码和页面引退	√	√
每个 VM 的多 GPU 数量	最多 16 个	最多 16 个

其他支持的 GPU

NVIDIA® Quadro RTX™ 6000、RTX 8000、NVIDIA P40、P100 以及用于刀片式服务器的 P6。

[1] Gartner. Market Guide for Server Virtualization（《服务器虚拟化市场指南》）. 2019 年 4 月 24 日. ID G00350674。
[2] 根据有效的“支持、更新和维护 (SUM)”合同提供。
[3] 不包含在 vComputeServer 许可证中，但可通过 NVIDIA NGC 支持服务合作伙伴单独获取。
[4] 适用于 Quadro RTX 6000 和 RTX 8000 的 24C 配置文件。
[5] 适用于 NVIDIA V100 的 32C 配置文件。
[6] 适用于 Quadro RTX 8000 的 48C 配置文件。