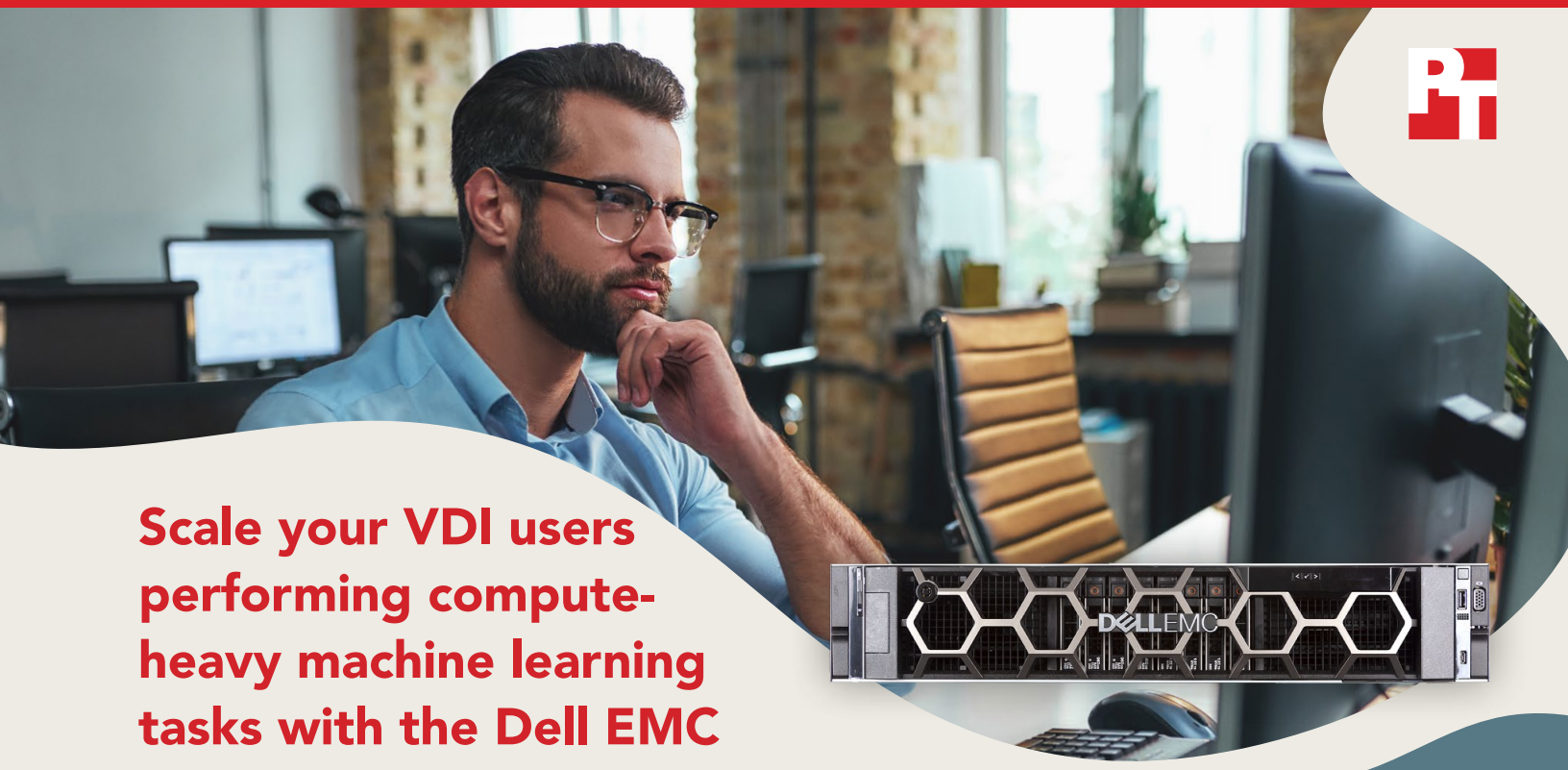


Scale your VDI users performing compute-heavy machine learning tasks with the Dell EMC PowerEdge R750xa

Featuring 3rd Generation Intel Xeon Scalable processors and up to four NVIDIA A100 40GB PCIe GPUs

For companies, tech start-ups, laboratories and others that work with machine learning and artificial intelligence (AI), GPU computing offers high performance. To support remote and campus-based employees logging into virtual desktops, you need a powerful VDI (virtual desktop infrastructure) solution. Dell EMC™ offers a flagship server aimed at GPU workloads. The Dell EMC PowerEdge™ R750xa supports two to four NVIDIA® A100 40GB PCIe GPUs per server—but which GPU configuration best suits your organization's needs?

At Principled Technologies, we assessed the VDI performance of a Dell EMC PowerEdge R750xa server featuring 3rd Gen Intel® Xeon® Scalable processors using two and four GPUs to run the AI performance benchmark MLPerf Inference v1.0 on VMware® Horizon® virtual desktops. When configured with four GPUs, the server supported up to 40 data science knowledge workers simultaneously performing computationally intense work; with two GPUs, the server supported 20 of the same type of worker. In both cases, the PowerEdge R750xa server supported the maximum number of users allowed to share each GPU while each user achieved a consistent level of work. Being able to comfortably support more users with high compute requirements can make for a better experience for each user connecting to your VDI network.

Scale up your VDI performance:

40 desktop users with four GPUs

20 desktop users with two GPUs

How we tested

Hardware

We tested a Dell EMC PowerEdge R750xa rack server in two configurations. The only difference between these configurations was the number of GPUs in the system (two or four). Table 1 highlights the hardware we used. Note that the PowerEdge R750xa supports the newest NVIDIA GPUs (such as A100, A30, and A10 GPUs) as well as Gen 3 GPUs (such as M10 and T4 GPUs).[1] For more detailed hardware information, see the Science behind this report.

Table 1: Hardware configuration highlights.

Server	Dell EMC PowerEdge R750xa
Processors	3rd Generation Intel Xeon Scalable processor (Intel Xeon Gold 6330)
Storage	4 x 960GB KIOXIA PM5 SAS SSDs 4 x 1.92TB Intel SSD D7 series PCIe 4.0 NVMe SSDs
Memory	32 x 64GB Hynix HMA with 3,200MHz memory
GPUs	2 or 4 x NVIDIA A100 Tensor Core (40GB PCIe)
Network adapters	1 x Broadcom Gigabit Ethernet BCM5720 1Gb dual-port NIC 1 x Intel Ethernet E810-XXV OCP 25Gb dual-port NIC

VDI workload

We used VMware View Planner to benchmark our VMware Horizon Version 2103 VDI environment on VMware vSphere 7 Update 2 and to populate the environment with simulated virtual users. To enable these different users to share GPU resources on a single server, we used NVIDIA vGPU software in conjunction with VMware to provide time-sliced vGPU instances to each virtual machine. We used "4c" vGPU instances, which allowed up to 10 vGPUs and, consequently, up to 10 VMs per GPU.

During our tests, we used VMware View Planner to emulate users signing into virtual desktops and completing simulated work. View Planner comes with several stock workloads; in addition to the "move_files" and "generate_random_files" workloads, we added a custom machine-learning workload which ran an Inference image classification test (MLPerf Inference v1.0 ResNet50).

Dell EMC PowerEdge R750xa

Dell considers the PowerEdge R750xa rack server their flagship server for GPU-based workloads. Powered by 3rd Generation Intel Xeon Scalable processors, the server contains the following features:[2]

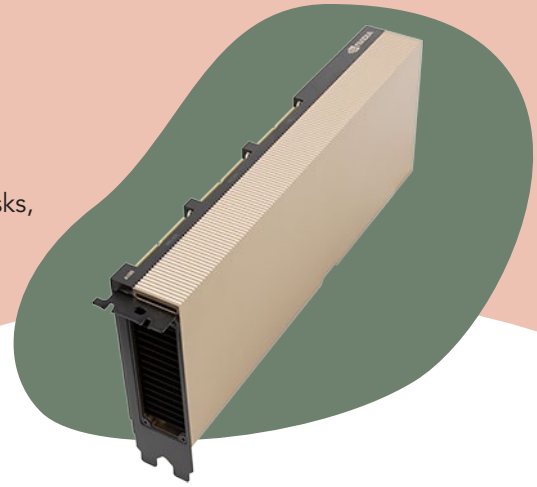
- Configurable with two to four GPUs
- Eight channels per CPU
- Up to 32 DDR4 DIMMS with a 3,200 megatransfers per second DIMM speed
- PCIe Gen 4
- Up to 8 SAS/SATA SSD or NVMe drives

To learn more, visit https://www.dell.com/en-us/work/shop/dell-poweredge-servers/poweredge-r750xa-rack-server/spd/poweredge-r750xa/pe_r750xa_14823_vi_vp.

NVIDIA A100 Tensor Core GPU (40GB PCIe)

NVIDIA positions its A100 Tensor Core GPU for artificial intelligence (AI), data analytics, and high-performance computing (HPC). According to NVIDIA, the GPU supports double, single, and half-precision compute tasks, unified virtual memory, and page migration engine.[3]

To learn more, visit https://www.nvidia.com/en-us/data-center/a100/.



Machine learning workload

MLPerf Inference is a benchmark suite that allows one to compare machine learning workload performance across machines. (To learn more, visit https://infohub.delltechnologies.com/p/introduction-to-mlperf-tm-inference-v1-0-performance-with-dell-emc-servers/.)

For our tests, we had each simulated VDI user complete the test Image Classification on Resnet50-v1.5. (For reference, image classification is the process of training a computer to correctly identify objects in photographs, and ResNet-50 is an inference model that organizations use for image classification.)

To represent a general GPU-limited task workload common among data science knowledge workers, we configured the MLPerf benchmark to occupy 100 percent of available GPU compute resources. We ran MLPerf in Server mode.

Note that MLPerf Inference does not represent the full spectrum of work that real-world data science knowledge workers would be responsible for. Rather, we aimed to sample the level of resource-intensive work that these types of users perform on a regular basis.

Note also that the MLPerf Inference ResNet-v1.5 results in this report have not been verified by the MLCommons Association.

What were we looking for?

For each server configuration, we wanted to determine how many simulated VDI users could run the MLPerf Inference workload at a minimum performance level (approximately 2,350 queries per second). To determine this minimum performance level, we ran the benchmark repeatedly with various settings, and settled on a configuration that produced a performance level large enough to ensure full GPU utilization without overloading the VMs.

VMware Horizon Version 2103

VMware Horizon is a virtual desktop infrastructure solution that enables IT admins to run applications on central data center hardware while allowing employees remote access to these applications as a service. According to VMware, Horizon version 2103 contains such improvements and new features as:[4]

- Instant clones
- Updates to Horizon Console
- Support for new REST API endpoints, PostgreSQL support for EventsDB, and federated access groups

To learn more, visit https://techzone.vmware.com/resource/what-vmware-horizon-and-how-does-it-work.

What we found

The PowerEdge R750xa supported 10 data science knowledge worker (DSKW) VDI sessions per GPU, which corresponds to the maximum number of virtual GPUs per physical GPU allowed by the NVIDIA vGPU driver. Importantly, in all our tests, each VDI user achieved consistent performance on a compute-intensive MLPerf Inference workload. Figure 1 compares the number of concurrent VDI users we were able to run on each configuration, with each user session maintaining a performance level of about 2,350 queries per second.

With four GPUs, the PowerEdge R750xa supported 40 data science knowledge worker users, which may make it a good option for organizations that need to support many workers who often perform compute-heavy work from their virtual desktops. Choosing to configure your servers with fewer GPUs may offer flexibility to an organization with a more varied mix of virtual desktop users.

Figures 2 and 3 show host resource data from each of the server configurations we tested. While the four-GPU configuration consumed 88 percent of CPU resources and 56 percent of memory resources, the configuration with two GPUs had lower usage—76 percent CPU and 28 percent memory. Purchasing servers with fewer GPUs would reduce hardware costs while making room for a variety of other users who use virtual desktops but don't need GPU computing.

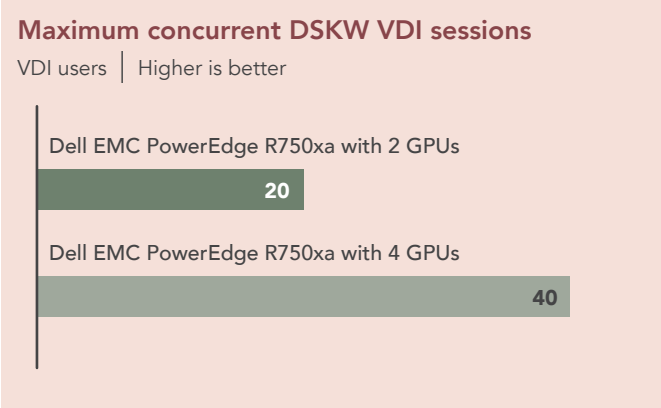


Figure 1: Maximum VDI sessions the solutions supported. Higher is better. Source: Principled Technologies

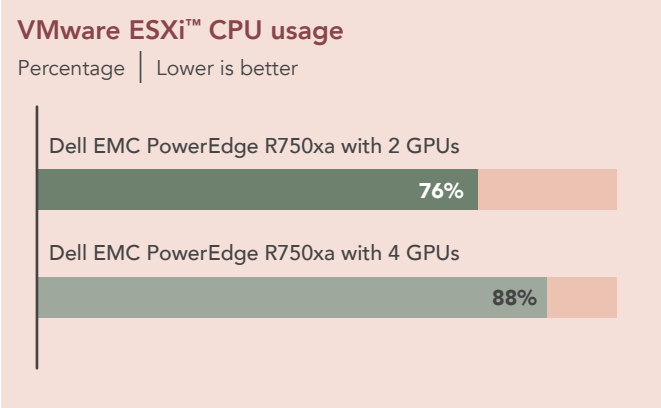


Figure 2: Processor utilization during VDI tests. Lower is better. Source: Principled Technologies

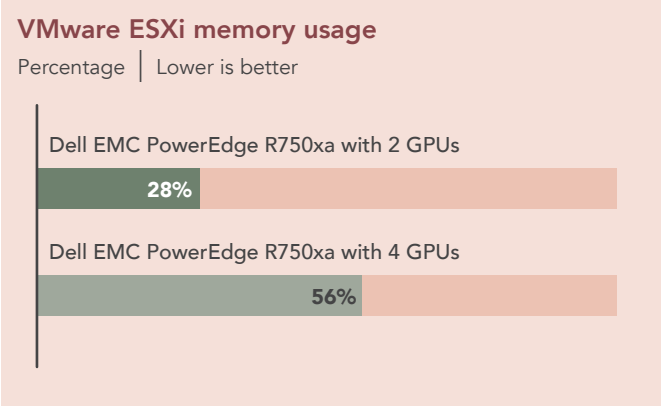


Figure 3: Memory usage during VDI tests. Lower is better. Source: Principled Technologies

Duty cycle, time-slicing, and capacity planning

Our tests show that the PowerEdge R750xa can support the maximum 10 users per GPU with each user commanding approximately 1/10th of the performance of a full GPU. However, in real-world scenarios, data science knowledge workers would not necessarily run the same compute jobs concurrently. With time-sliced scheduling of NVIDIA vGPU, a single user could use nearly 100 percent of the GPU compute cycles when no other users are running jobs that consume GPU. The user would still be limited in the amount of allocated vGPU memory; however, an admin can configure the NVIDIA vGPU driver to fairly schedule compute cycles among contending users just as we did during testing.

If your data science workers require fixed performance for high-duty-cycle workloads (such as a web-facing inference server), you may use NVIDIA vGPU in "MIG" mode, which partitions one physical GPU into as many as seven physical vGPUs.

Conclusion

GPU computing is an important resource for an organization's data science knowledge workers. With the rising popularity of virtual work environments, your organization may be interested in adopting GPU computing in a VDI environment.

In our tests, we found that a Dell EMC PowerEdge R750xa with 3rd Gen Intel Xeon Scalable processors could support 10 virtual users per NVIDIA A100 40GB PCIe GPU (the maximum the NVIDIA drivers support) while each user completed a compute-intensive machine learning workload (MLPerf Inference - Image Classification on ResNet-50 v1.5)

When configured with four GPUs, the PowerEdge R750xa supported 40 concurrent VDI users. In our two-GPU tests, the PowerEdge R750xa supported 20 users while saving CPU and memory resources that an organization could reserve for users with fewer compute resource needs. Based on our test results, organizations that support a mix of data science knowledge workers and workers with fewer compute resource needs may consider configuring PowerEdge R750xa servers with fewer GPUs, while organizations that need the infrastructure to keep up with many data science knowledge workers may want to configure their servers with the maximum number of GPUs to receive the greatest benefit.

---

1 "Configuration details," accessed August 10, 2021, https://infohub.delltechnologies.com/l/mlperf-tm-inference-v1-0-nvidia-gpu-based-benchmarks-on-dell-emc-poweredge-r750xa-servers/configuration-details-44..

2 Dell EMC PowerEdge R750xa spec sheet," accessed July 20, 2021, https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-R750xa-spec-sheet.pdf.om_last/.

3 NVIDIA A100 Tensor Core GPU," accessed July 20, 2021, https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/A100-PCIE-Prduct-Brief.pdf.

4 Cindy Carrol, "What's New in VMware Horizon Version 8, App Volumes, and Dynamic Environment Manager (2103)," accessed July 20, 2021, https://techzone.vmware.com/blog/whats-new-vmware-horizon-version-8-app-volumes-and-dynamic-environment-manager-2103.

Read the science behind this report at http://facts.pt/DCxN6al ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.