A Principled Technologies report: Hands-on testing. Real-world results.

Achieve near-bare-metal inference throughput for image classification workloads with the Dell PowerEdge R7525 server using virtual GPUs

Plus, virtualizing servers with VMware vSphere 7.0 Update 3 and vGPUs could help simplify machine learning workload management

Virtualization has changed the world of computing, but every hypervisor adds some amount of computing overhead compared to a bare-metal environment. While this overhead may not present an issue for many workloads, businesses may well feel those effects on compute-heavy workloads, such as machine learning (ML). With VMware® vSphere® 7.0 Update 3, Dell™ PowerEdge™ R7525 servers powered by AMD EPYC™ processors offer a platform that can minimize performance loss with virtualized GPUs (vGPUs) while delivering the ease-of-management, redundancy, reliability, mobility, and security advantages that virtualization can provide.

In our data center, we tested ML-based image classification inference throughput on bare-metal and virtual configurations of a Dell PowerEdge R7525 with two AMD EPYC 7543 processors and an NVIDIA® A100 Tensor Core GPU. When running the MLPerf™ ResNet50 workload, the bare-metal configuration processed 28,130 queries per second. The virtual configuration, with VMware vSphere 7.0 Update 3 and a single guest VM, processed 27,435 queries per second—97.5 percent of the bare-metal configuration's image classification performance. Either configuration of the PowerEdge R7525 could serve as the backbone in image-classification workloads, delivering strong performance as a bare-metal or virtual solution. Virtualizing the PowerEdge R7525 with NVIDIA vGPU software and vSphere 7 Update 3 could enable IT admins to run multiple virtual machines (VMs) on the same physical GPU, making it easier to dial in GPU-level performance as needed.

Minimize performance degradation from virtualization

Achieve 97.5% of the image classification performance of a bare-metal configuration

Simplify ML workload management

Admins can ensure uptime and resource allocation through virtual management

How we tested

Artificial intelligence (AI) and ML workloads rely heavily on computing speed. Organizations that want faster AI or ML workload performance can use GPUs in addition to CPUs because GPUs offer extreme parallelization and matrix math capabilities. Organizations may run GPU-based workloads on bare-metal configurations to avoid the possibility of hypervisor or virtualization overheads. For those compute-heavy workloads, IT staff configure optimization, tuning, and power settings to provide the best GPU-level performance possible.

For our bare-metal solution, we configured a Dell PowerEdge R7525 server with Ubuntu 20.04. The PowerEdge server contained 480 GB of storage and 512 GB of DDR4 3200MT/S memory. We installed, tuned, and ran an MLPerf AI benchmark on the bare-metal solution according to best practices. The workload we chose classifies images within a known dataset using the ResNet50 ML model.

We built our virtual environment on the same server hardware as our bare-metal environment. After testing the bare-metal solution, we kept the same BIOS settings, installed VMware vSphere 7.0 Update 3 onto the server, and created a VM with the same OS and tunings as our bare-metal configuration. Then, we ran the same MLPerf ResNet50 workload on our VM and compared the results of the two configurations.

Although our testing used only one VM, supporting more VMs is possible. By making vGPU software available in vSphere 7.0 Update 3, many VMs can share one or more GPUs. Prior to NVIDIA vGPU (and predecessor NVIDIA GRID vGPU™), GPUs were available to VMs only in pass-through mode, meaning a VM using a GPU would lock it down from use by other VMs. Sharing GPU resources between many VMs could allow your organization to use one GPU for multiple tasks. For example, one Linux® VM could run an AI or ML workload while a Windows VM runs a graphic-intensive ANSYS application for computer-aided engineering (CAE). Before vGPUs, running multiple workloads on the same GPU was complex and inefficient, much less doing the same for multiple VMs.

Potential real-world benefits

Organizations using AI and ML may run training and inference workloads on separate servers because the workloads run at different times and with different power requirements, tunings, or access restrictions. If a server goes down, its workload could be offline until it is brought up on a new server or IT fixes the issue that brought the server down.

For manufacturing companies that run image classification workloads as part of quality control, that downtime could lead to lost revenue and delays in production. With Dell PowerEdge R7525 servers supporting VMware vSphere 7.0 Update 3 virtual environments, such a company could combine pairs or multiples of servers into redundant clusters, with workloads separated by VM. Then, if a server goes down, IT staff can quickly bring up the VMs on a failover server, with little interruption or downtime.

Virtual environments need, at most, as much hardware as equivalent bare-metal environments. However, it's possible that a virtual environment with VMware vSphere 7 and NVIDIA vGPU software could handle the same amount of work as a bare-metal environment with less hardware or do more work with more hardware.

Classify images quickly with a bare-metal Dell PowerEdge R7525 environment

While this study highlights the benefits of GPU virtualization, it is worth noting that the Dell PowerEdge R7525 equipped with AMD EPYC processors and an NVIDIA A100 Tensor Core GPU delivered solid bare-metal performance for the image classification workload we tested. To enable the performance that the NVIDIA A100 Tensor Core GPU provided, the Dell PowerEdge R7525 supported Gen 4 PCIe connectivity, with up to 16 GT/s, to help maximize GPU throughput and offered 1400W dual PSUs and an optional GPU-enablement kit with high-powered fans, heatsinks, and air shroud to help keep the GPU cool. In addition, the AMD EPYC CPUs provided ample processing power for supporting tasks, such as copying between memory and GPU and I/O logic. Having a powerful server as a host for the NVIDIA A100 Tensor Core GPU provided the necessary system resources, power, and cooling to help deliver smooth performance during our tests. Our bare-metal solution processed an average of 28,130.7 queries per second on the MLPerf ResNet50 workload.

Bare-metal GPU performance on image classification workload

28,130 **Average queries per second (higher is better)**

Figure 1: The average number of queries processed per second on the bare-metal Dell PowerEdge R7525 environment. Source: Principled Technologies.

About the Dell PowerEdge R7525

The Dell PowerEdge R7525 server, powered by 3rd Gen AMD EPYC 7003 series processors, can provide PCIe™ Gen 4 speed with up to three double-width 300W or six single-width 75W accelerators. Dual 2400W PSUs provide the necessary power to support GPUs, and up to 12 hot-plug fans can keep graphics cards cool.[1]

To learn more about the PowerEdge R7525, visit https://www.delltechnologies.com/en-us/servers/poweredge-rack-servers.htm.

Virtualize a Dell PowerEdge R7525 server running VMware vSphere 7.0 Update 3 and get comparable image classification performance

On a VM in our vSphere 7.0 Update 3 environment using vGPUs, the MLPerf ResNet50 workload processed an average of 27,435.1 queries per second, which was 97.5 percent of the queries per second that the bare-metal solution classified (see Figure 2).

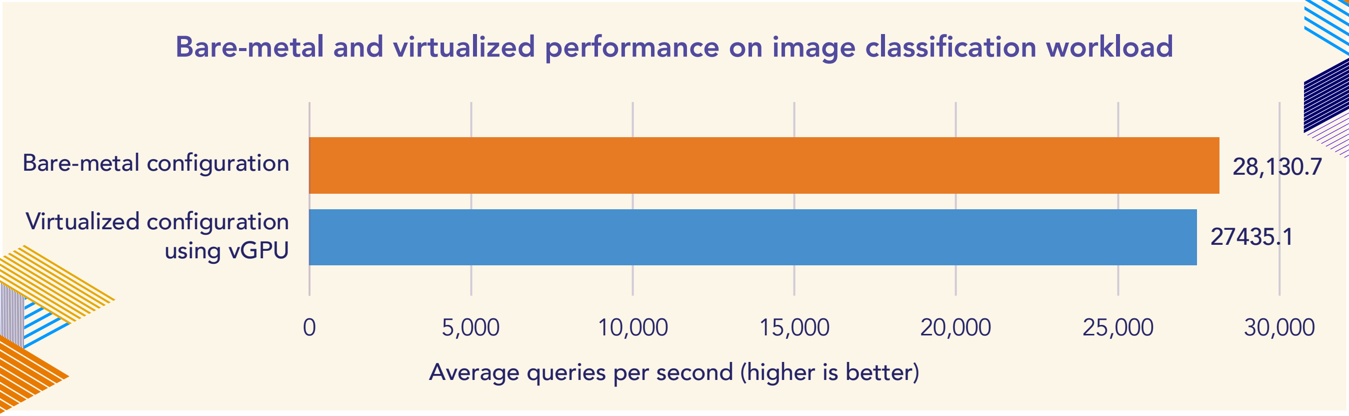

Figure 2: The average number of queries processed per second on the bare-metal and virtualized Dell PowerEdge R7525 environments. Source: Principled Technologies.

In addition to the comparable image classification performance in our testing, Dell PowerEdge R7525 servers with VMware vSphere 7.0 Update 3 can allow admins to migrate, back up, resize, and perform all the other benefits of virtualization on GPU-enabled VMs. This includes multiple VMs running on a single host, shared virtual GPUs, cloning, vMotion™, distributed resource scheduling, and suspending/resuming VMs.

About VMware vSphere 7.0 Update 3

Released in Q3 2021, vSphere 7.0 Update 3 aims to help organizations running and developing AI and ML workloads. The release supports the latest generation of GPUs from NVIDIA, NVIDIA Virtual GPU (vGPU) software,[2] NVIDIA GPUDirect RDMA for vGPUs,[3] and the latest spatial partitioning-based NVIDIA multi-instance GPUs.[4] The software also offers many capabilities for vSphere with Tanzu.[5]

To learn more about vSphere 7.0 Update 3, visit https://www.vmware.com/products/vsphere.html.

Conclusion

For Dell PowerEdge R7525 servers with AMD EPYC processors and NVIDIA A100 Tensor Core GPUs, running VMware vSphere 7 Update 3 and NVIDIA vGPUs allows you to share GPU cores for individual VMs to run ML workloads. In our image classification test, a virtual configuration of the PowerEdge R7525 server with AMD EPYC processors and NVIDIA vGPUs achieved up to 97.5 percent of bare-metal performance on the same server. This means that your organization could get the benefits of virtualization, such as flexibility, reliability, and ease of management, without incurring a significant performance hit on image classification workloads.

[1] Dell Technologies, "Dell EMC PowerEdge R7525 Technical Guide," accessed May 27, 2022, https://i.dell.com/sites/csdocuments/Product_Docs/en/PowerEdge-R7525-Spec-Sheet.pdf.

[2] VMware, "vSphere 7 Update 3 - What's New Technical Overview," accessed June 1, 2022, https://www.youtube.com/watch?v=yD3cvKpKwZ4.

[3] NVIDIA, "Virtual GPU Software R384 for VMware vSphere Release Notes," accessed June 1, 2022, https://docs.nvidia.com/grid/5.0/grid-vgpu-release-notes-vmware-vsphere/index.html.

[4] NVIDIA, "NVIDIA Multi-Instance GPU and NVIDIA Virtual Compute Server," accessed June 1, 2022, https://www.nvidia.com/content/dam/en-zz/Solutions/design-visualization/solutions/resources/documents1/Technical-Brief-Multi-Instance-GPU-NVIDIA-Virtual-Compute-Server.pdf.

[5] VMware, "Announcing vSphere 7 Update 3," accessed June 1, 2022, https://blogs.vmware.com/vsphere/2021/09/announcing-vsphere-7-update-3.html.

Read the science behind this report at https://facts.pt/lmXpGA7 ▶

This project was commissioned by Dell Technologies.

Principled Technologies®

Facts matter.®