Processed a total of

29,896 samples per second
with 10 containers sharing 1 GPU

34,352 samples per second
with a single container using 1 whole GPU

Combine containerization and GPU acceleration on VMware: Dell PowerEdge R750 servers with NVIDIA GPUs and VMware vSphere with Tanzu

Our results running a vGPU-accelerated deep learning image-classification workload in this environment

More and more enterprises are turning to artificial intelligence (AI), including machine learning, with two-thirds of respondents to a 2021 McKinsey survey saying that their AI investments will grow over the next three years.[1] For organizations that use machine learning workloads for critical business operations, VMware® vSphere with Tanzu containerized environments provide the ability to rapidly scale and robustly deploy applications. The most computationally demanding workloads require containers with GPUs, which can process multiple computations simultaneously, making them particularly well-suited to artificial intelligence and deep learning applications. Virtualizing GPU hardware allows multiple containers to share GPU resources when appropriate for the workload.

At Principled Technologies, we used ResNet-50—a deep learning image classification workload—on a Dell™ PowerEdge™ R750 server with an NVIDIA A100 Tensor Core GPU running VMware vSphere® with Tanzu. With 10 containers sharing the GPU, the PowerEdge R750 server with NVIDIA GPU processed up to 29,896 samples per second. With a single container using all of the GPU resources and a larger batch size, performance increased to a maximum of 34,352 samples per second.

VMware vSphere with Tanzu + Dell PowerEdge R750 server with NVIDIA A100 Tensor Core GPU = a robust platform for containerized machine learning workloads

Our testing brought together several components: Dell PowerEdge servers, Kubernetes containerization on the VMware Tanzu platform, and NVIDIA virtual graphics processing units (vGPUs).

To run their applications, many companies use containers, which pull together everything an application needs to run. Kubernetes is an open-source platform for deploying and managing applications that run in containerized environments.

NVIDIA vGPU software "creates virtual GPUs that can be shared across multiple virtual machines, accessed by any device, anywhere."[2] Companies can use NVIDIA vGPU software for a wide range of workloads. This approach combines the management and security benefits of virtualization with the performance of GPUs, from which many modern workloads can benefit.

We set out to see how well a Dell PowerEdge R750 equipped with an NVIDIA A100 Tensor Core GPU in a cluster with two Dell PowerEdge R7525 servers could handle a VMware vSphere and VMware Tanzu Kubernetes environment for a machine learning app. Because we needed a multi-node cluster to use VMware vSphere with Tanzu, we created a cluster with this server and two Dell PowerEdge R7525 servers without GPUs. (In a real-world setting, the PowerEdge R7525 servers would be available to run non-GPU workloads. In our testing, they were idle.) We used the ResNet-50 deep learning inference workload, which classifies images and measures machine learning capabilities and generates a metric of samples per second.

We ran a series of tests to determine the maximum number of samples per second the ResNet-50 image-classification workload could process across multiple instances with different numbers of containers and different batch sizes. We ran a variety of configurations of vSphere with Tanzu worker-node VMs and concurrent ResNet-50 instances to explore performance in terms of throughput. We used one vGPU and one container per VM. This process reflects the work an IT administrator might perform to determine the optimal configuration for hosting any similar image-classification workload on a comparable cluster. As the utilization of GPU resources on the NVIDIA A100 Tensor Core GPU neared 100 percent, we determined that we had reached the optimum number of vSphere with Tanzu worker nodes and ResNet-50 instances and achieved our maximum samples-per-second rate for that number of containers. (We also experimented with various tuning parameters to confirm this throughput rate was maximal.)

About ResNet-50

Resnet-50 is a model that organizations use for image classification, or the process of correctly identifying objects in images. Strong image classification performance may be vital for use cases in safety and security, retail, healthcare, manufacturing, and other markets. In our testing, we used the Resnet-50 Offline implementation from the MLPerf inference benchmark suite, which uses the Resnet-50 model and measures the number of (inference) samples per second a solution processes.

Our test environment

The test environment we used was as follows:

- Server nodes
  - One Dell PowerEdge R750 server (running VMware vSphere with Tanzu Kubernetes)
    - Intel® Xeon® Gold 6330 processor
    - 1 TiB RAM
    - NVIDIA A100 Tensor Core GPU
  - Two Dell PowerEdge R7525 servers (idle in our testing)
    - AMD EPYC 7763 processors
    - Because using VMware vSphere with Tanzu required a multi-node cluster, we created a cluster with these servers and the Dell PowerEdge R750 server.
- VMware vSphere 7.0 Update 3 with VMware Tanzu Kubernetes environment
  - Tanzu 1.6
  - Kubernetes 1.21.6
  - Ubuntu® 20.04 with EFI enabled
  - We installed Tanzu on all three server nodes, but the Tanzu workload cluster used only the PowerEdge R750 with GPU, which hosted all of the VMs and containers.
- NVIDIA AI Enterprise version 2.0
  - NVIDIA Grid vGPU driver version: 510.47.03

About the Dell PowerEdge R750 servers

The Dell PowerEdge R750 is a full-featured, general purpose 2U rack server powered by 3rd Gen Intel Xeon Scalable processors. According to Dell, the PowerEdge R750 is purpose-built to optimize application performance and acceleration with PCIe Gen 4.0 compatibility, eight channels of memory per CPU, and up to 24 NVMe™ drives. For workloads that benefit from GPUs, the servers support up to two double-width 300W accelerators, three single-width 150W accelerators, or six single-width 75W accelerators.[3] It also includes "improved air-cooling features and optional Direct Liquid Cooling to support increasing power and thermal requirements."[4]

To learn more about the features that the Dell PowerEdge R750 offers, visit https://www.dell.com/en-us/ work/shop/productdetailstxn/poweredge-r750.

What we learned

To gain an understanding of the ResNet-50 performance possible in our test environment, we tested both the minimum number of containers possible, one, and the greatest number, 10. This is based on the number of vGPU slices possible from the NVIDIA A100 Tensor Core GPU because you can subdivide each GPU into a maximum of 10 slices. We assigned one vGPU slice per container.

We performed each test with custom provisioning and orchestration scripts using Ansible, Python, and bash to install and run MLPerf Inference v2.1 codes. We deployed Tanzu clusters with one worker node hosting one container for each vGPU slice allowed for both vGPU slice types we tested, so one node and container for the 40GiB slice, and 10 nodes and containers for the 4GiB slice. We gave each node 16 GiB of memory per 4 GiB of vGPU memory, or 160 GiB for the 40GiB slice and 16 GiB for the 4GiB slice. All nodes had 16 vCPUs. Table 1 presents some of the key parameters and results of our testing. For complete details on testing, see the science behind the report.

Table 1: Key parameters and results of ResNet-50 image classification performance testing with different container counts. Higher throughput is better. Source: Principled Technologies.

Test parameters			Test results (samples per second)	
vGPU RAM (GiB)	Tanzu node count (number of containers)	Batch size (number of images per batch)	Per-node throughput	Aggregate throughput
40	1	2,048	34,352	34,352
4	10	128	2,989	29,896

About VMware vSphere with Tanzu

VMware vSphere with Tanzu is "the new generation of vSphere for containerized applications."[5] According to VMware, vSphere with Tanzu allows IT administrators to use their "existing vSphere environment to manage multiple clusters alongside virtual machines through vCenter, delivering Kubernetes clusters at a rapid pace."[6]

VMware vSphere with Tanzu facilitates AI/ML workloads by supporting compute accelerators such as NVIDIA GPUs. With supported GPU virtualization or pass through, it's possible be flexibly provision compute resources to AI/ML workloads. According to VMware, "vSphere enables developers, system administrators, and DevOps teams to accelerate all types of AI and ML workloads, whether VM-based or containerized, by leveraging the full power and performance of the latest generation of NVIDIA GPUs."[7]

Learn more at https://www.vmware.com/products/vsphere/vsphere-with-tanzu.html.

As Figure 1 shows, the maximum samples-per-second rate the server achieved with a single container was approximately 15 percent greater than with 10 containers. We can speculate that this increased performance was due to the larger batch size, which leads to more efficient use of the GPU. GPU memory limits batch size, so a larger vGPU permits larger batch sizes. VM virtualization overhead and GPU virtualization overhead also affect performance. The number of containers and batch size you employ will depend on your specific use case. Smaller or more sporadic jobs benefit from the flexibility of many smaller vGPU slices, while larger or more regular jobs benefit from the dedicated memory and compute of a whole GPU.

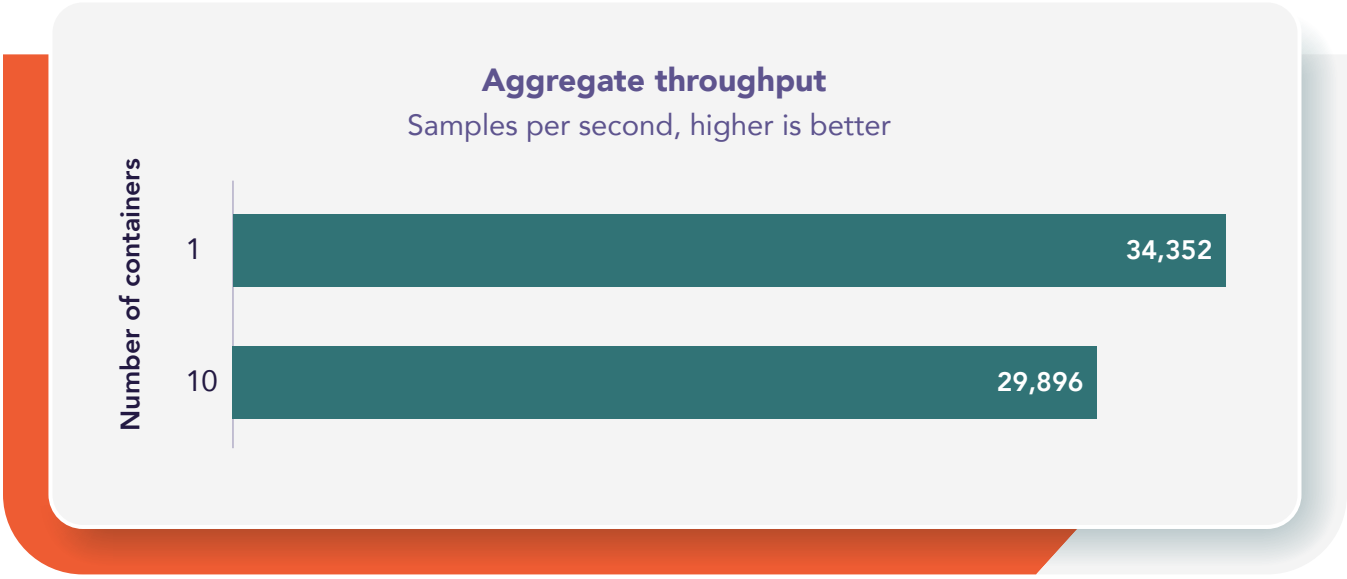

Figure 1: ResNet-50 image classification performance, in samples per second, achieved by the Dell PowerEdge R750 with an NVIDIA A100 Tensor Core GPU in a VMware vSphere 7.0 Update 3 with VMware Tanzu environment with different container counts. Higher is better. Source: Principled Technologies.

About NVIDIA A100 Tensor Core GPU

According to NVIDIA, the A100 Tensor Core GPU delivers "unprecedented acceleration at every scale to power the world's highest-performing elastic data centers for AI, data analytics, and HPC."[8] To adjust to fluctuating demands, you can partition the A100 into up to ten vGPUs or up to seven MIG-mode GPUs.

NVIDIA AI enterprise integrates with VMware vSphere with Tanzu to provide virtualization of NVIDIA GPUs—including the A100—to both virtual machines and containers. You can deploy Tanzu clusters with vGPU-enabled worker nodes by installing system drivers, deploying a license server, creating a machine class with desired vGPU, and deploying the NVAIE Kubernetes operator.

Learn more about the NVIDIA A100 Tensor Core GPU at https://www.nvidia.com/content/dam/en-zz/Solutions/Data-Center/a100/pdf/nvidia-a100-datasheet-nvidia-us-2188504-web.pdf.

Conclusion

We ran a deep learning image-classification workload on a Dell PowerEdge R750 server with an NVIDIA A100 Tensor Core GPU that was in a VMware vSphere with Tanzu Kubernetes container environment along with two other servers. Using GPU virtualization to allow 10 containers to share the single GPU, our test solution achieved a maximum of 29,896 samples per second. With a single container using all of the GPU resources and a larger batch size, the solution achieved a maximum of 34,352 samples per second. These results show that the PowerEdge R750 with an NVIDIA A100 Tensor Core GPU in a VMware Kubernetes environment with GPU virtualization can support flexibly apportioning GPU compute capability across multiple machine learning workloads in Kubernetes clusters.

1. "The state of AI in 2021," accessed March 2, 2023, https://www.mckinsey.com/business-functions/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021.
2. NVIDIA, "Unlock Next Level Performance with virtual GPUs," accessed March 2, 2023, https://www.nvidia.com/en-us/data-center/virtual-solutions/.
3. Dell, "Dell EMC PowerEdge R750 Spec Sheet," accessed March 2, 2023, https://i.dell.com/sites/csdocuments/Product_Docs/en/poweredge-R750-spec-sheet.pdf.
4. Dell, "Dell EMC PowerEdge R750 Spec Sheet"
5. VMware, "vSphere with Tanzu," accessed March 2, 2023, https://www.vmware.com/products/vsphere/vsphere-with-tanzu.html.
6. VMware, "vSphere with Tanzu."
7. VMware, "vSphere AI/ML Solutions," accessed March 2, 2023, https://www.vmware.com/products/vsphere/ai-ml.html.
8. NVIDIA, "NVIDIA A100 TENSOR CORE GPU," accessed March 2, 2023, https://www.nvidia.com/en-us/data-center/a100/.

Read the science behind this report at https://facts.pt/BPtAk4o ▶

Principled Technologies®

Facts matter.®

This project was commissioned by Dell Technologies.