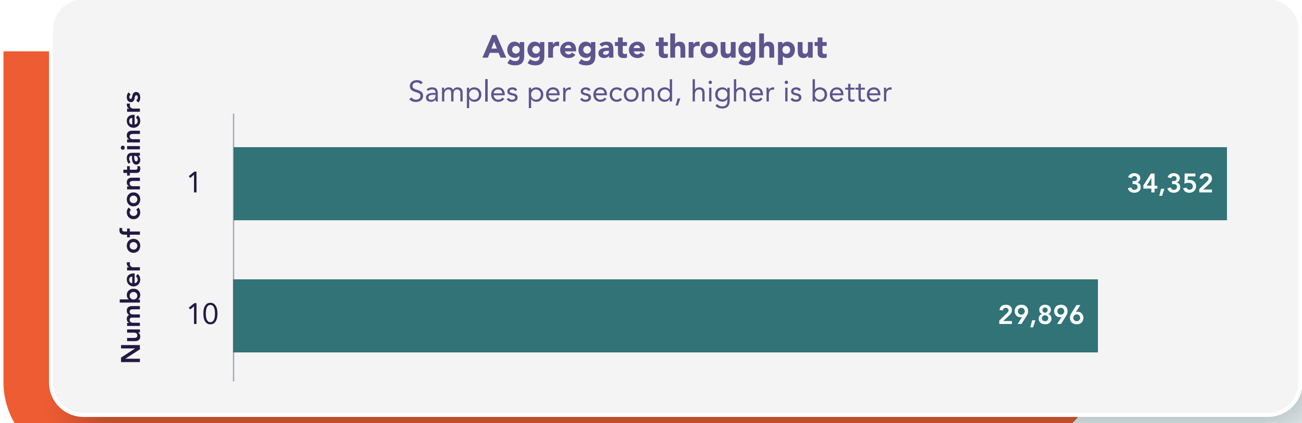
Combine containerization and GPU acceleration on VMware: Dell PowerEdge R750 servers with NVIDIA GPUs and VMware vSphere with Tanzu

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

We used a ResNet-50 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 an NVIDIA GPU processed up to **29,896 samples per second**.

With a **single container** using all the GPU resources and a larger batch size, performance increased to a maximum of **34,352 samples per second**.



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

You can determine the optimal configuration for hosting any similar image-classification workload on a comparable cluster:

- Smaller or more sporadic jobs benefit from the flexibility of many smaller vGPU slices while delivering nearly the same overall performance as using the whole GPU
- Larger or more regular jobs benefit from the dedicated memory and compute of a whole GPU

Learn more at https://facts.pt/Hi5jvB2

Principled Technologies®