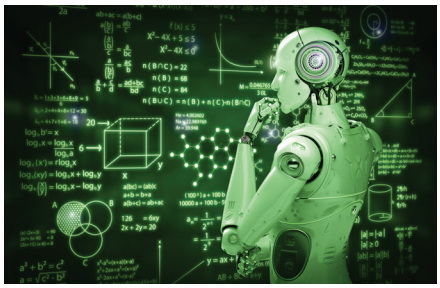
SOLUTION BRIEF

Agile Infrastructure at Cloud Scale for Machine Learning

The DriveScale Composable Platform

Machine learning (ML) is used to advance a broad range of business and customer capabilities. Designing the infrastructure required for machine learning is complex with each part to the process requiring different server infrastructure. Statistical machine learning requires a smaller data set and can be run on a standard server. Compute and data-intensive deep neural networks require much larger data sets and high-performance technologies such as GPUs and low-latency network fabric. Production deployments, also called inference, are less computationally intensive with lower requirements for compute and storage yet may still require low-latency.



To control for performance, latency, data gravity, and cost, most companies are deploying ML in their own data centers or co-location facilities. The challenge is building out or expanding ML infrastructure, while reducing complexity, increasing productivity and getting the performance your solution needs.

DriveScale offers a unique approach to compute and data infrastructure for machine learning. With DriveScale, you can create your own compute-intensive or data-intensive server platform out of software on-the-fly. Enabling users to create the exact server and storage configuration needed for the workload, the DriveScale Composable Platform leverages disaggregated low-cost compute, storage and high-performance Ethernet fabric. With DriveScale software, you can create high scale, high performance, heterogeneous solutions to meet the exact requirements of each ML workload.

DriveScale Delivers Agile Server Infrastructure for Machine Learning

The DriveScale Composable Platform is the only server infrastructure that scales and adapts compute and storage resources to meet the needs of applications on the fly. Heterogeneous, low-cost compute nodes and GPU nodes can be composed as part of the server platform along with 100G connected NVMe flash drives or 10G connected hard disk drives in

DriveScale

SOLUTION BRIEF

Agile Infrastructure at Cloud Scale for Machine Learning

dense eBODs (Ethernet Box of Drives). With DriveScale, users can deploy server and storage infrastructure in minutes not months, maximize resource utilization and eliminate wasted spend with independent compute and storage upgrades. In addition, DriveScale enables one solution to deploy ML infrastructure for all needs from data ingest to training to inference to archival using the optimal resources to fit the workload.

DriveScale provides NVMe over Flash using RoCEv2 to enable a high performance, low-latency Ethernet-based solution which accelerates random small file I/O operations and minimizes bottlenecks common with the ML training workload. In addition, with DriveScale, flash drives can be carved into slices as small as 1GB and attached to individual compute nodes or GPU nodes to ensure optimal utilization.

With DriveScale, you choose your preferred vendors for diskless CPU-centric and GPU-centric servers, and for Ethernet-attached hard disk and flash drives, and then easily compose heterogeneous compute and storage configurations to meet the needs of individual workloads. You can add or remove compute or storage resources as needed or replace failed compute or storage in seconds from the DriveScale platform.

Why DriveScale for Machine Learning

The DriveScale Composable Platform enables IT to compose server infrastructure from a software application combining the flexibility and agility of cloud with the performance and latency requirements of machine learning.

The DriveScale platform provides the ability to:

- Create high-performance all-flash configurations
- Carve flash drives into slices for optimal utilization
- Connect 1 or 100s of drives to CPU/GPU nodes
- Maintain low latency and direct-attached storage performance
- Optimize utilization of GPUs, CPUs and drives
- Recover instantly from compute or drive failures from an easy-to-use software interface

By optimizing resource utilization using DriveScale, companies can deploy server infrastructure at a lower cost than alternatives while ensuring they have the flexibility to quickly scale up or down compute and storage resources as needed.

Composable Infrastructure

Composable Infrastructure is next-generation server infrastructure that provides the ability to flexibly create, adapt, deploy and later redeploy servers using pools of disaggregated, heterogeneous compute, storage and network fabric. According to IDC, the composable infrastructure market is estimated to grow from $752 million in 2018 to $4.7 billion in 2023.