VEXATA SOLUTIONS FOR AI AND MACHINE LEARNING

ENABLING PREDICTIVE AND COGNITIVE ANALYTICS

Machine Learning (ML) workloads are increasing in volume and complexity as organizations look to reduce training and operational timelines for artificial intelligence (AI) use cases. This has given rise to massively parallel GPU servers like the Nvidia DGX-1, delivering massive compute power to run these machine learning frameworks.

In order to accelerate training and operational cycles, storage systems that power these AI/ML pipelines must maintain ultra-low latency, massive ingest bandwidth and heavy mixed random and sequential read/write handling. Architectures using direct attached storage (DAS) limits performance and data mobility, while existing all-flash arrays lack the sustained performance to deliver timely insights at scale. Only Vexata can deliver the performance and scale of NVMe to accelerate ML workloads.

AI/ML USE CASES

- Fraud Analytics
- Quant Trading
- Industrial IoT
- Computer Vision
- Speech Recognition
- Hyper Spectrometry
- Biomedical Cancer Detection

Nvidia DGX-1+ VX-100FS File Storage System

ACCELERATE ML/AI WORKLOADS WITH VEXATA

Vexata VX-100FS, with its transformative VX-OS is purpose built to overcome these machine learning challenges

Reduce training and inferencing time from days to hours, improving data scientist productivity

- Accelerated data path with deterministic low latency performance for better GPU utilization
- Native NVMe solid state performance delivered with standard NVMe-oF (RoCE) and file system interfaces

Access large training and inferencing data-sets

- Accelerated non-blocking access to NVMe media for large data ingest with low latency IO performance.

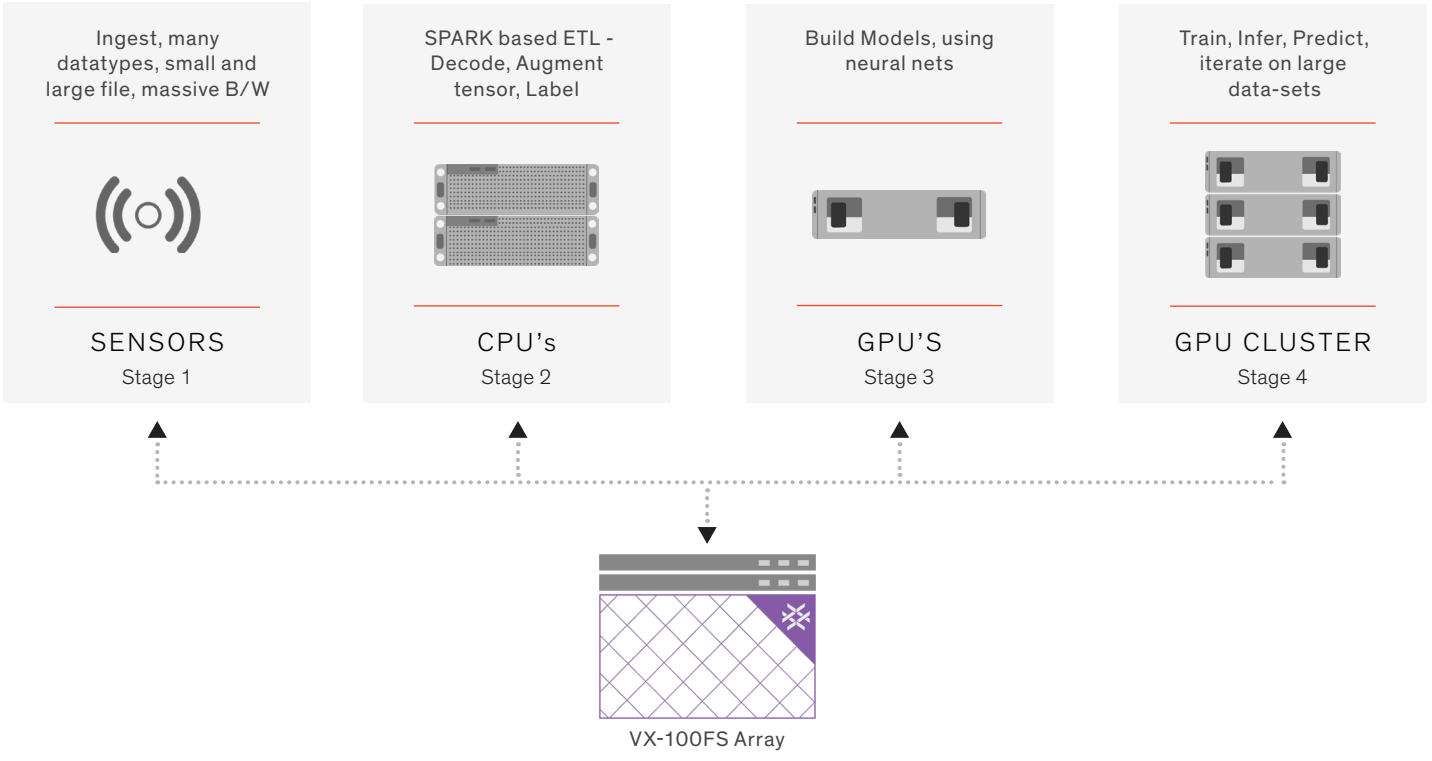
Consolidate and eliminate movement between data pipeline stages

- Shared storage to handle all data pipeline stages without performance degradation
- Simultaneously supports small block random IO , large block sequential IO, mixed Read/Write IO
- In-place data analytics with flexibility of ingest protocols (FC, NVMe-oF, NFS, SMB, S3)

Enterprise storage services, security and data protection

- Instantaneous snapshots and clones, replication, RAID 5/6 protection to eliminate 3 copy replication, compression and 256bit encryption

Accelerated Data Pipeline



VEXATA SOLUTIONS FOR AI AND MACHINE LEARNING

CUSTOMER CASE STUDY

Customer Challenges

- Advanced Machine Learning/AI pipeline for Computer Vision, Hyper-Spectrometry for cancer detection with requirements for massive data ingest and multi-stage training and inference.
- Originally deployed a DAS storage architecture forcing staging to local SSD's, resulted in very long neural net training cycles (20+ days), wasted GPU cycles ($$) and Data Scientist time ($$$)

Attempted to deploy a first generation all-flash file storage solution, but technology had insufficient bandwidth and poor small file performance characteristics

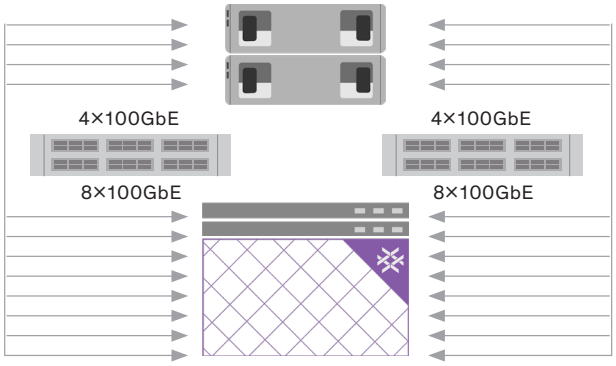
Customer Benefits with Vexata

- Faster time to results and accurate prediction with Vexata
- Lower total cost of ownership, due to pipeline stage consolidation and data scientist utilization
- Better utilization of Nvidia DGX-1 servers

5 TIMES BETTER THAN COMPETITION

Reference Architecture

- Two DGX-1 systems (8 Tesla V100 GPU's, 2× Intel E5-2698 v4)
- 2 PFLOPS of Deep Learning performance
- Container based Nvidia GPU Cloud Deep Learning stack with machine learning frameworks
- Mellanox SN2100 100GbE switch (2 switches)
- Vexata VX-100FS NVMe-oF scale-out storage system
- Up to 435TB of fast file storage tier
- 50 GB/s of bandwidth, 1M NFS IOPS
- Scale – Add DGX's, add head nodes, add arrays



	LEADING FLASH BLADE VENDOR 4 DGX SERVERS, 15 BLADES			VEXATA Nvidia SOLUTION - 2 DGX SERVERS, 8 BLADES, 2 HEADS			VEXATA Nvidia SOLUTION - 4 DGX SERVERS, 16 BLADES, 4 HEADS		
FILE SIZE	Available B/W - training/ inference	Images/sec	Remaining Bandwidth	Available B/W - training/ inference	Images/sec	Remaining Bandwidth	Available B/W - training/ inference	Images/sec	Remaining Bandwidth
150KB	5 GB/s	33K	5GB/s	12.5 GB/s	83K	12.5 GB/s	25 GB/s	166K	25 GB/s
1MB	7.5 GB/s	7.5K	7.5 GB/s	12.5 GB/s	12.5K	12.5 GB/s	25 GB/s	25K	25 GB/s

TEST CONFIGURATION:

- Bandwidth equally divided between training/Inferencing and Ingest/ETL/Build
- Imagenet pre-trained model – Alexnet used because it is storage IO heavy
- Inception V3, Resnet – 50, Resent – 152, Alexnet, VGG16 container images
- Supervised Learning, labelled images, 1.28M, 1000 categories
- Standard docker file - nvcr.io/nvidia/tensorflow:18.04-py2
- Batch_size = 64

ABOUT VEXATA: Vexata is the leader in active data management solutions. Vexata's unique breakthrough enterprise offerings enable transformative performance and scale from database and analytics applications. With unparalleled ability to consume the latest in media like NVMe Flash and now with Intel Optane™ SSDs, Vexata systems deploy simply and seamlessly into existing storage environments. Learn more at www.vexata.com

Contact Vexata: info@vexata.com

SB-1020-06072018