VMware Cloud Foundation and vSAN Storage for Artificial Intelligence

Silverton Consulting, Inc. StorInt™ Briefing

Introduction

Organizations today are struggling to integrate Artificial Intelligence (AI) into their work activities. Some challenges to adoption include lack of expertise, clean/filtered training data, and insufficient hardware/software infrastructure.

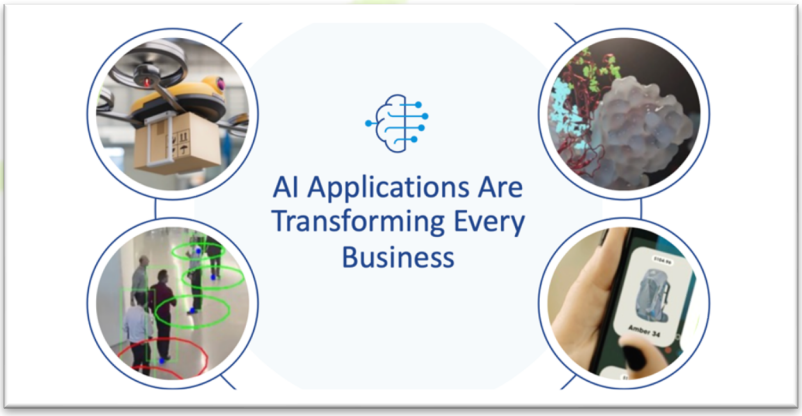
While expertise will improve over time, training data and infrastructure will remain obstacles to AI adoption. Effectively cleaning and filtering massive amounts of customer and operational data for use as training data for AI will continue to be difficult.

As for infrastructure, storage for data access, AI stack functionality and GPU access are all major concerns. Besides the limited availability of GPUs, running GPUs continues to be cost prohibitive. These factors make it even more important to have storage that can utilize GPUs as much as possible during training and other AI activities. Software support for full stack AI adoption is also difficult to select, configure, and use.

Storage performance and software are critical. As we will discuss in this paper, VMware vSAN™ storage, together with VMware Cloud Foundation™ AI workloads, can greatly facilitate enterprise AI adoption and deployment.

AI data requirements

The real benefit of AI adoption lies in deploying trained models that use AI inferencing to draw conclusions from data. But before inferencing can happen, AI models must be trained and this needs TBs to exabytes of cleaned, filtered, and tagged data.



AI data for model training uses file or object storage. Structured data can also play a role in model training, typically as extracts to files or objects.

During model training, file and object data are read in multiple times in random order to train complex **deep neural networks**. This process continues until model results are properly calibrated.

Model training is both a random read IO and GPU compute-intensive activity. During AI model training, write IO is used occasionally to copy or checkpoint neural networks and log training activity.

Silverton Consulting
Strategy, Storage & Systems

Traditionally, AI models are trained to perform such functions as converting audio to text, identifying image objects, classifying text sentiment, and identifying customers needing assistance. And for Generative AI Large Language Models (GenAI LLMs), create text, code, or images from prompts.

AI model deep neural networks are often categorized by their number of parameters, which can range anywhere from 1M to 1B for enterprise-class AI models and to a trillion or more for GenAI LLMs.

Training for enterprise-class models requires a modest number of GPUs and servers and relatively little time. For example,

- **ResNet 50**, an image classification model (~25.6M parameters), can be trained with two NVIDIA A100-PCIe-80GB GPUs in a matter of hours.[1]
- **BERT-large**, a natural language processing model (~345M parameters), can be trained with six NVIDIA L4 GPUs in a matter of hours.
- **Stable Diffusion v2,** a text to image GenAI model (~890M parameters), can be trained with eight NVIDIA H100-SXM4-80GB GPUs in under an hour.

Older GPUs will train models more slowly and newer GPUs will train models more quickly. Also, the more GPUs used, the faster a model can usually be trained.

One main concern with GenAI LLMs is that they can hallucinate answers to prompts, creating nonsensical or inaccurate outputs because they perceive patterns or objects that do not exist. Another concern is that they may reveal proprietary information. To facilitate better and more compliant responses to GenAI prompts, organizations often utilize the following:

- **LLM fine-tuning**, which is a GenAI LLM "mini"-training run that adds an organization's information to help it generate better responses.
- **Retrieval-augmented generation (RAG)**, which loads corporate information into a vector or RAG DB that can then be searched with prompt text. The results can be used to supplement prompts with additional contextual information for response generation.

Both LLM fine-tuning and RAG can reduce hallucination and deliver more focused GenAI LLM results.

Inferencing uses a trained AI model in applications – it's essentially an AI model in action. Inferencing may use GPUs for low-latency applications, but it can also use CPU compute alone. In either case, inferencing takes orders of magnitude less compute power than model training.

[1] All training results listed here are from MLCommons, MLPerf Data Center Training benchmarks, as of 19 Mar 2024. See: https://mlcommons.org/benchmarks/training/.

There are three types of inferencing:

- **Batch or offline inferencing**, where multiple requests are read from a file or object, which is then supplied to a model for inferencing. Batch inferencing is typically throughput sensitive. A key metric for batch inferencing is queries per second (Q/s).
- **API, server, or online inferencing**, where a website or online application transaction calls a model to perform an inference. API inferencing can be both throughput and latency sensitive. Key metrics for server inferencing are Q/s and transaction latency.
- **Stream inferencing**, where a model analyzes a video, sensor or message/transaction stream and tries to detect anomalies or other items of interest. Stream inferencing is mostly latency sensitive. The key metric for stream inferencing is latency.

IO activity for inferencing can range from very high-performance-throughput intensive to more normal, mixed IO. For example,

- A ball detection model that highlights ball activity on a field, analyzes a video stream arriving at high frame rates. The model may require very low-latency response to highlight the ball in each frame. In this case, the video stream is throughput, read-write intensive, and inferencing is latency sensitive.
- A temperature sensor stream from a set of jet engines uses much less data. Inferencing to determine engine anomalies may or may not require as fast a turnaround, but this work is unlikely to require high performance IO.

For any AI inferencing, model inputs and output inferences should be recorded for compliance, governance, model drift detection (data change over time) and new training data.

Inferencing can happen in the data center core, in co-location facilities or in the cloud. If latency is a concern, inferencing may also occur at the edge. Storage to support inferencing must support the performance and deployment requirements dictated by the AI application.

AI storage needs

The storage infrastructure needed to support enterprise model training and inferencing must be able to handle the heavy random read-intensive IO needed for training, as well as the diverse IO performance and deployment requirements for inferencing.

As described above, most enterprises can perform GenAI LLM fine-tuning or RAG. Moreover, training enterprise-class AI models in the small (<1M parameters) to medium (1M to 1B parameters) size range can use only modest quantities of GPUs and servers. To train these models, storage must support file and object random, read-intensive access with occasional

write activity for model checkpointing and logging. Data used to train enterprise-class AI models are often under a few TBs in size.

Batch and server inferencing IO varies considerably depending on the application. However, it's most likely to be mixed read-write, file or object IO. As mentioned earlier, some inferencing may need to be done at the edge to support very low latency.

VMware vSAN Express Storage Architecture

VMware vSAN Express Storage Architecture (ESA) is a new generation of vSAN storage. It builds on the previous generation vSAN Original Storage Architecture (OSA), which has been powering VMware clusters for more than 10 years. This latest version of vSAN adds important new capabilities that improve IO performance, increase storage capacity, and enhance usability.

VMware vSAN ESA only supports NVMe SSDs for storage. Besides the faster IO from NVMe SSDs, vSAN ESA fault domains contain only a single physical NVMe SSD to increase availability.

In addition, vSAN ESA has adopted a new log structured file system approach that changes how data are written. This approach significantly reduces IO overhead, resulting in faster IO for RAID protection. Further, vSAN ESA added Adaptive RAID-5, which allows RAID-5 data protection to be used with three hosts in edge deployments. And for enterprises using data-at-rest security, vSAN ESA offers lower latency IO by only performing encryption once at the top of the stack.

All these IO performance enhancements allow vSAN ESA clusters to support up to 500 VMs per host versus 200 VMs for vSAN OSA.

VMware vSAN offers Storage Policy Based Management (SPBM) to simplify the use and management of storage by the enterprise. Using SPBM, AI and other workloads can be guaranteed the resources required to maintain performance and other policy objectives. With SPBM, customers can establish requirements for workloads and users and it will automatically ensure those requirements are maintained on future storage requests.

Further, vSAN ESA provides more real-time monitoring of storage performance, and it includes new PowerCLI integration, improved performance bottleneck detection and key management, as well as simpler device management and Skyline Health enhancements.

VMware vSAN ESA supports **stretched clusters** for high availability, which replicate data from one cluster to the other and perform cross-site sharing of storage. In fact, vSAN ESA offers up to 20 ESXi hosts at each side of a stretched cluster serving storage and VM execution.

With vSAN ESA, data mirrored between sites is compressed before transmission, reducing bandwidth requirements, and providing higher write throughput over data links. Furthermore, vSAN ESA coalesces small writes into bigger blocks before writing to SSDs and across to remote sites. This action should reduce SSD IOs and data link activity and thus improve stretched cluster IO performance for small, block-intensive IO.

In addition, VMware vSAN ESA supports remote storage where one vSAN cluster can share storage with other vSAN clusters. Normally, nodes in a vSAN cluster have similar storage- and compute-sized hosts across a cluster, which can be a limitation for customers with diverse storage and compute requirements. However, remote storage allows separate vSAN clusters to have different compute-to-storage ratios and customers can access storage on either cluster depending on application requirements. As we discuss in the next section, VMware **vSAN Max** takes this configuration flexibility to another level.

VMware vSAN ESA supports the **Data Persistence platform (DPp)**, which supplies S3-compatible object storage access over vSAN storage. Enterprises should keep DPp in mind for AI model training. As we discuss below, given vSAN ESA's IO performance, ESA storage should perform better for DPp deployment.

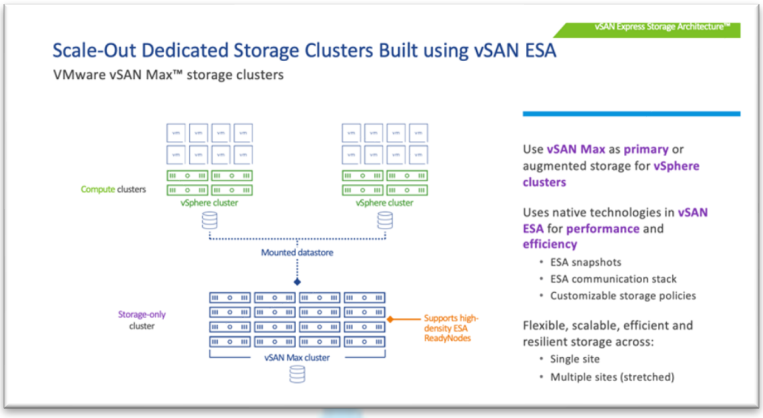
Moreover, VMware vSAN ESA supports the **vSAN Distributed File Service (vDFS)**, which provides NFS and SMB file storage over vSAN storage. Again, as we discuss later, the IO performance of vSAN ESA should result in better performing vDFS.

The protocol flexibility, IO performance and data capabilities of vSAN ESA can readily support most inferencing activity done at the core, in co-location facilities, in the cloud or at the edge. Moreover, vSAN ESA can easily support small (<1M parameters) AI model training and possibly support medium-sized (1M to 1B parameters) model training, depending on hardware configurations and model requirements.

vSAN Max on vSAN ESA

VMware's vSAN Max on vSAN ESA offers a fully disaggregated storage solution for enterprises. Unlike base vSAN ESA, vSAN Max customers can create storage-intensive clusters that don't

offer VM compute. Like remote storage, vSAN Max clusters offer shared storage services to other clusters in the data center.



With vSAN Max, compute and storage can scale entirely independently. Some vSAN Max advantages include:

- **Storage and compute-only clusters** – customers can configure a cluster of storage nodes that only support storage services with no VM compute or compute-only clusters that only support VM execution with no storage.
- **Storage-dense nodes** – vSAN ESA limited storage configurations to support concurrent VM execution; however, customers can use vSAN Max to deploy as many storage drives as slots available in servers to support storage-dense clusters.
- **Storage capacity scalability to multi-PB or more** – with non-vSAN Max clusters, storage capacities were restricted to what could be deployed in an HCI configuration. With vSAN Max clusters and storage-dense servers, such limitations are significantly reduced.
- **Higher performing storage** – with storage-only clusters, servers can be configured to support higher IOPS and bandwidth because vSAN Max clusters don't need to be optimized for both data IO and VM execution.

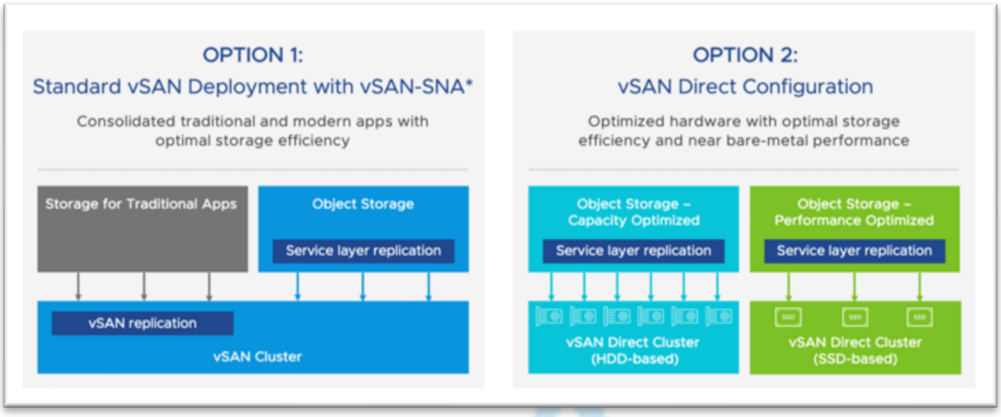
All the vSAN ESA IO performance improvements discussed earlier can further boost read and write IO performance and data bandwidth for vSAN Max clusters. In addition, vSAN Max storage-only clusters give compute node servers even more room to hold GPUs in support of AI work.

vSAN DPp object store

As mentioned earlier, VMware vSAN DPp supports object storage access on vSAN storage. The DPp can operate under vSAN ESA or vSAN Max clusters. Using vSAN Max on vSAN ESA offers higher scalability and better IO performance, making it an optimal deployment for DPp services in support of AI.

VMware vSAN DPp is delivered using software partners that include DELL™, MinIO™ and Cloudian™. DPp provides native object storage access to applications that need it. For instance, AI model training can use DPp to gain direct access to object data hosted on vSAN.

For DPp, vSAN offers two modes of operation: **vSAN Shared Nothing Architecture (SNA)**, which uses normal vSAN data protection services for objects, and **vSAN Direct** without vSAN data protection, which provides a direct path to vSAN storage for DPp IO.



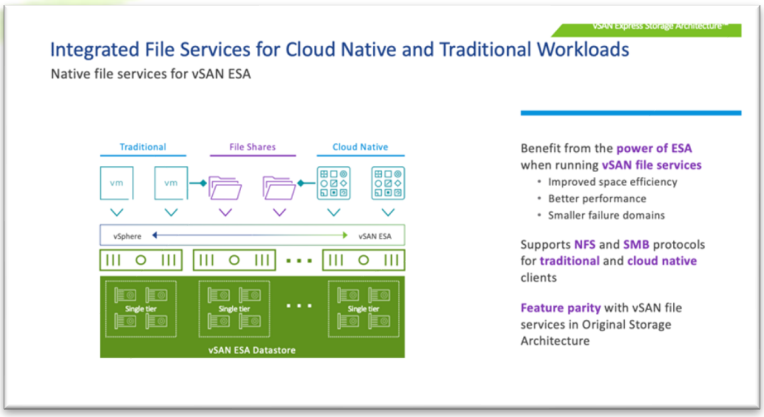
Whenever the object storage application itself provides data protection or where no protection is needed, using vSAN Direct over SNA reduces IO overhead and thus increases IO performance to read and write object data.

DPp runs as a container or cloud-native application on VMware. Multiple DPp services can operate over a single vSAN cluster. Different DPp deployments can also use vSAN SNA or vSAN Direct depending on application requirements.

As mentioned, object storage is very popular for AI model training. In fact, much of the data used to train standard models are available in S3 object storage format. Many RAG deployments also use objects for backend storage and can be hosted on top of DPp storage.

vSAN Distributed File Service

VMware vDFS provides NFS and SMB file access for vSAN storage. VMware vDFS can operate on vSAN ESA, as well as on vSAN Max clusters.



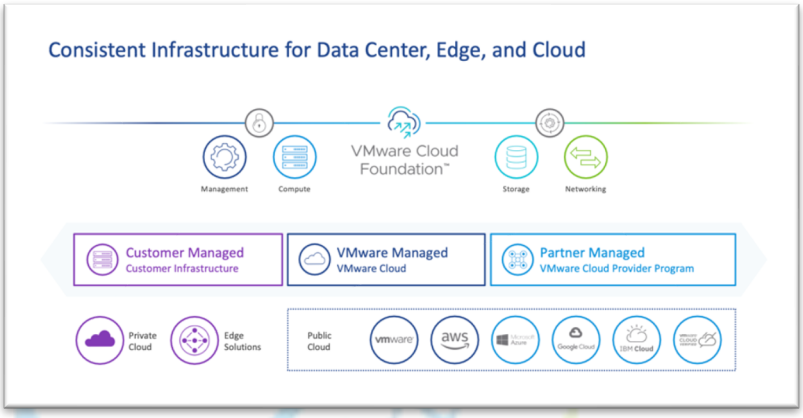
With vDFS, AI training and other applications that need NFS file access can deploy file data on vSAN and access it using native file services.

Because vDFS can operate over any vSAN cluster, customers can use vSAN Max to configure their vDFS storage for AI training with high capacity and high random read performance. Alternatively, customers can use vSAN ESA to configure their vDFS storage for AI inferencing with more mixed IO capabilities.

Silverton Consulting
Strategy, Storage & Systems

VMware Cloud Foundation

VMware Cloud Foundation (VCF) provides VMware vSAN clusters that operate across multiple environments in the public cloud, in the private cloud, in co-location centers, on prem and even at the edge. VCF supports a single control plane to manage multiple VCF instances, with each instance potentially offering multiple vSAN clusters.



VCF supports cloud-native (container) applications with options for more sophisticated cloud-native management. Recall that vSAN DPp is a cloud-native application. Moreover, VCF K8s™ clusters can be deployed and spread across VM pods within a vSAN cluster to accelerate application development and deployment.

In addition, VCF supplies high levels of NVIDIA™ integration, which include sharing NVIDIA GPUs across VMs as well as support for the full **NVIDIA AI Enterprise** stack.

VMware recently introduced **VMware Private AI Foundation.** Customers can use Private AI to gain more control over the information used for LLM generation. VMware Private AI Foundation also provides an open-source AI stack reference model, which can be used to deploy open source LLMs with corporate fine-tuning or RAG.

VCF offers NVIDIA integration, multiple deployment options and support for vSAN ESA, vSAN Max, DPp and vDFS. Customers with AI training and inferencing in the core, in the cloud or at the edge should consider running their AI work on VCF.

Full stack AI software

Another significant challenge with any new workload is how to decide the right software to use to implement it. With AI's high interest and expanding ecosystem, the AI stack is undergoing rapid change and innovation, with several layers with each layer having many options. One would like to think that options available for each layer would be interchangeable. Sadly, you would be wrong. As such, each option will need to be integrated with all the other options across an AI stack.

VCF AI can simplify the work required to integrate an AI stack. As discussed earlier, VCF offers both NVDIA AI Enterprise stack and VMware Private AI Foundation, two full stack alternatives to ease AI stack choice and speedup AI adoption.

Findings and Recommendations

For enterprise customers deploying AI workloads in their environments, we recommend vSAN ESA, that can provide an optimal storage platform for an enterprise's current and future workloads and an overall great platform for AI inferencing activities in the enterprise. In addition,

- The capacity scalability and higher IO performance of vSAN Max storage enables it to sustain the high random read IOP performance for traditional enterprise AI model training and LLM fine-tuning/RAG, as well as the high read-write IO required for inferencing.
- vSAN DPp would be ideal storage for any customer that needs object data to perform AI model training, LLM fine-tuning or RAG at the core or in the cloud. With the option of hosting DPp on vSAN Max clusters, object data storage capacity and IO performance shouldn't be limitations.

Summary

Enterprise adoption of AI can prove challenging. While there's much more to AI than just storage and software functionality, IO performance often becomes a significant AI training or inferencing bottleneck. Moreover, building your own AI software stack with its complex set of options and layers, all of which need to operate well together, to deliver AI functionality needed for the enterprise, can be a significant and ongoing challenge.

Data capacities and storage IO requirements for AI differ substantially depending on whether an enterprise is training gigantic LLM models, fine-tuning LLMs, using RAG, training more realistic enterprise-class AI models or simply performing inferencing.

VMware vSAN ESA, vSAN Max and vSAN DPP can supply all the IO performance and storage capacity needed to train enterprise class AI models as well as the performance required to sustain inferencing in the cloud, data center and at the edge.

Finally, while LLM training may initially require special-purpose infrastructure, more modest infrastructure suffices for fine-tuning LLMs, and RAG use. With all the interest in deploying LLMs in the enterprise, it makes good business sense for enterprises to deploy enterprise GenAI applications on VMware's VCF using vSAN storage.