Enterprise Strategy Group | Getting to the bigger truth.™

ESG WHITE PAPER

Enabling an AI-ready Infrastructure with VMware

Enterprise Agility, Reliability, and Efficiency with a Flexible Infrastructure Powered by HCI and GPUs

By Paul Nashawaty, Senior Analyst; and Mike Leone, Senior Analyst

February 2022

This ESG White Paper was commissioned by VMware
and is distributed under license from TechTarget, Inc.

Contents

Overview

Enterprise organizations are looking to embrace artificial intelligence (AI). In fact, in the past 12-18 months, many organizations are already well on their way down the AI path and most are struggling with how to operationalize it at production scale.[1] They want to benefit from the game-changing and transformative benefits that can be realized from AI. A key component of embracing AI is arming data science teams with tools that enable timely success in achieving their AI goals. Data science teams want to leverage their preferred frameworks and learning algorithms through their preferred applications and tools. They are looking to embrace automated workflows and tasks from data selection and training to deployment. And throughout those different steps of the AI lifecycle, ensuring right-sized resources are available is critical.

So, what does this mean to organizations? Organizations must work to identify and understand the key AI use cases that matter the most to their business, establish an efficient data pipeline, and deliver a performant, reliable infrastructure. To get there, agility is essential—being able to quickly build, iterate, and deploy AI-based applications at scale and handle the real-time requirements of the business. All of this is enabled by utilizing a foundational infrastructure architecture that offers the flexibility and scalability demanded by the potential diverse and dynamic AI workloads, from experimentation to training to eventually deploying models into production. And when it comes to the underlying architecture, IT teams must ensure the consistent delivery of high-performance storage, powerful compute via GPUs that can be shared across teams and workloads, and automation and orchestration of the delivery of said resources fueled by self-service.

Challenges in Delivering an Effective AI Environment

While systemic challenges remain ever-present across most organizations, infrastructure readiness is proving to be a challenge. Between skills gaps throughout the AI lifecycle, weak links throughout the infrastructure stack, and aggressive timelines, organizations appear to be more set up for AI failure than AI success.

Skills Gaps

As organizations look to ramp up AI initiatives, the first roadblock comes in the form of existing skills gaps. In fact, ESG research shows that more than 1 in 3 (36%) organizations have a problematic IT skills shortage in AI.[2] This is particularly troublesome with IT often outright owning the final decision for AI infrastructure purchases. This is not to say that organizations don't have AI IT experts, but even those AI experts that are in organizations are often overburdened as they are asked to incorporate new technologies like storage-class memory (SCM) or GPUs into existing infrastructures. The fact of the matter is that those existing infrastructures often lack capabilities that align to the advanced requirements of key AI workloads. For IT, this equates to constant troubleshooting, management complexity, and support tickets, especially when working with infrastructure that isn't aligned to IT standards. With the introduction of AI-ready infrastructure, organizations are in the early stages of seeing the potential of tying in hyperconverged infrastructure with virtualization and automation to better enable IT teams to respond to the demands of AI stakeholders and workloads.

Infrastructure Shortcomings

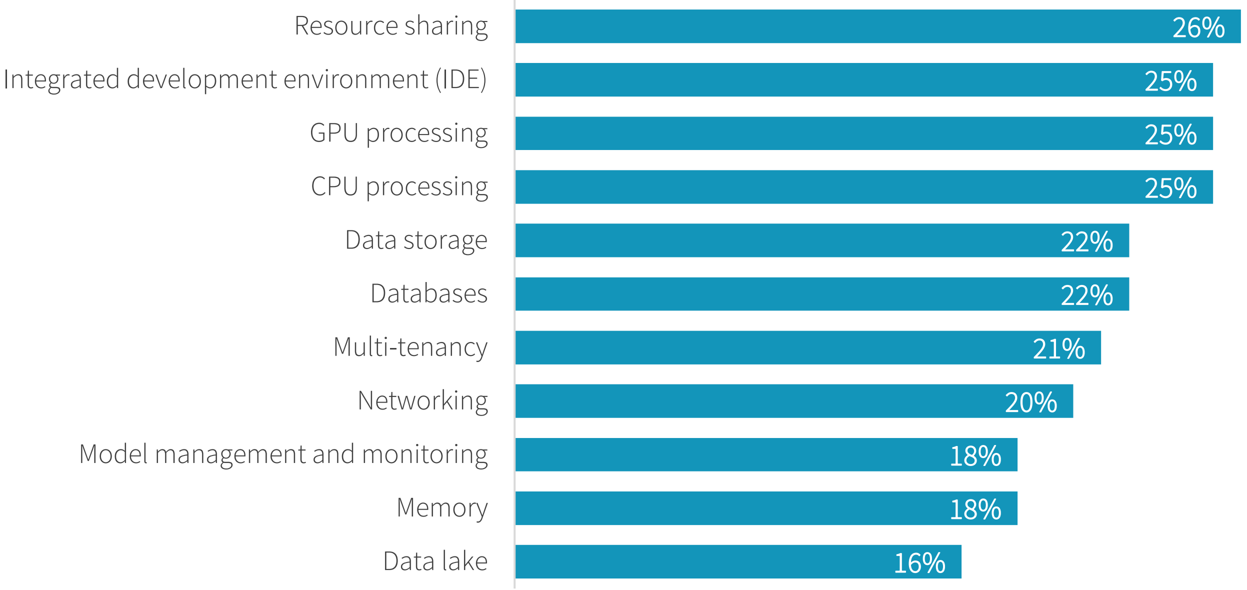
AI infrastructure requires optimization and tight integration of several components and resources across both software and hardware. In other words, simply adding a GPU or two to an existing infrastructure deployment is not yielding the results IT, data scientists, or line of business teams desire. Existing technology (and previous investments) simply cannot keep up with the performance and concurrency demands of AI workloads. Between inadequate processing power, storage capacity, and networking capabilities, along with an inability to properly manage resource allocation, infrastructure

[1] Source: ESG Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.
[2] Source: ESG Complete Survey Results, *2022 Technology Spending Intentions*, November 2021.

readiness is proving to be a significant issue. While infrastructure cost remains a top challenge for organizations as they look to embrace AI, according to ESG research, 26% of surveyed organizations reported that they believed resource sharing was or would be among the top weakest links in their ability to deliver an effective AI/ML environment, followed by an integrated development environment (25%), processing (both CPU and GPU – 25% each), and data storage (22%).[3]

Figure 1. Weakest Links in the AI Infrastructure Stack

Which parts of the infrastructure stack do you believe are or will be the weakest links in your organization's ability to deliver an effective AI/ML environment? (Percent of respondents, N=325, three responses accepted)



Source: ESG, a division of TechTarget, Inc.

Further, the dynamic workloads found through the AI lifecycle, from ingest and data preparation to training and inference, have different requirements. General-purpose components stitched together in a DIY implementation or even as-a-service are proving to fall short, especially as diverse AI workloads require more customization in how hardware and software are deployed and how resources are provided to stakeholders to support their customized workloads. That means a need to constantly configure and reconfigure systems based on the workload in order to optimize performance and scale.

Lifecycle Bottlenecks

Availability of an optimized system, with the tools, technologies and underlying data creates several bottlenecks, all of which impact the time to value of AI initiatives. When pressed on the areas of the AI lifecycle that cause the most headaches, it was not data science-specific tasks that landed at the top, but areas at the beginning and end of the cyclical AI lifecycle. Organizations often get lost in data science aspects of AI (i.e., model building, feature engineering, training, etc.), as these tend to create headaches and complexity. The fact of the matter is that other areas of the AI lifecycle are more of a concern and should be prioritized. In fact, according to ESG research, 37% of respondents indicated that deployment is one of the phases of the AI lifecycle that causes them the greatest headaches, 41% cited data transformation/preparation/wrangling, and 34% selected data integration.[4]

[3] Source: ESG Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.
[4] Ibid.

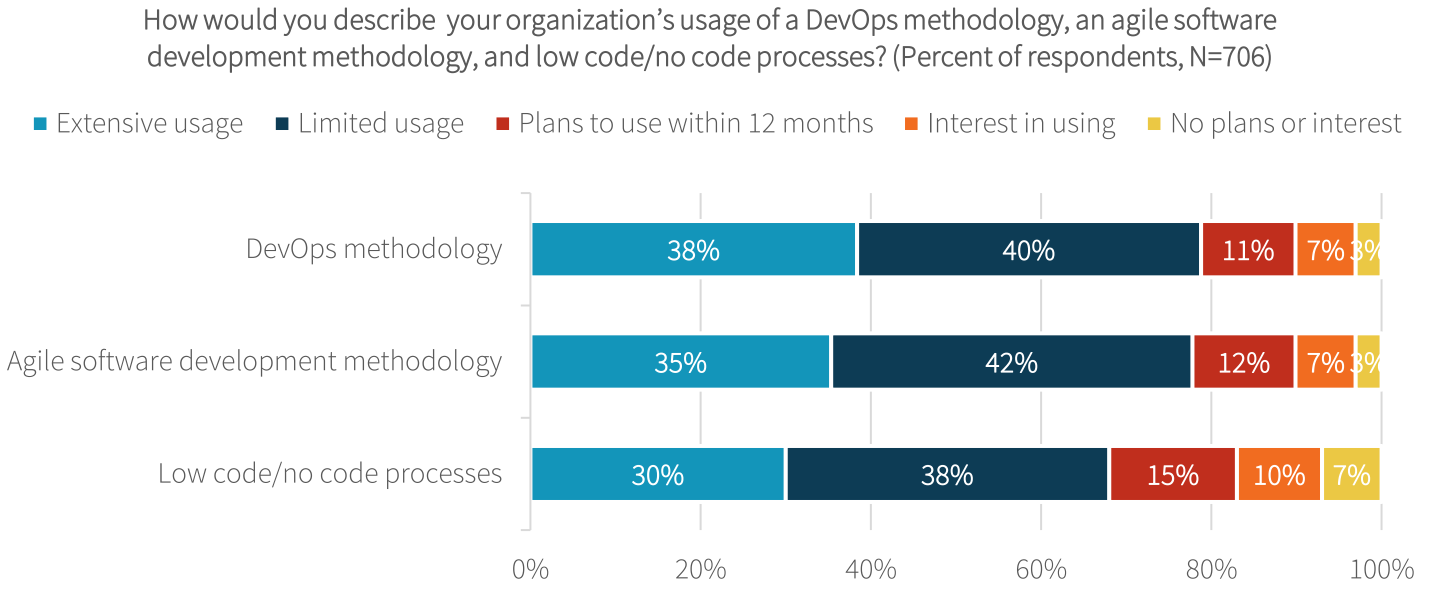
In addition, organizations need to be able to account for the different number of tools that must be tightly integrated and compatible, including versions, libraries, and drivers. They also cannot overlook the undertaking of moving from a test environment with sample data and complete control to a large-scale distributed environment across on-premises and public cloud environments.

Put all these challenges together, and IT teams are being driven to look for full-stack solutions that enable them to standardize on the same technologies, simplify resource allocation and management, and deliver right-sized infrastructure based on AI workload demand, from experimentation and testing, to training and tuning, and eventually deploying models for inferencing. Having a robust set of APIs that allows the dev teams to self-serve in an automated fashion while maintaining the IT guardrails for security/compliance is important for organizations to manage their own resources and delivery. The ability to configure/reconfigure clusters to meet the dynamic requirements of the data pipeline is an area of focus and should be considered with deployments.

The Push for Simplicity and Agility

Organizations are pushed to look for ways to simplify the development and deployment of AI processes via increased agility empowered by DevOps. For AI workloads, the challenges of reconfiguring infrastructure and critical components such as GPUs need to be optimized for tasks like workload placement and mobility. This requires an agile methodology to rapidly deliver against the business needs. DevOps methodologies are ways to address consistency and deliver rapid results from development to deployment. In ESG's recent *2022 Technology Spending Intentions Survey*, respondents indicated broad adoption when it comes to leveraging DevOps methodologies, agile development practices, and no-code/low-code processes (see Figure 2).[5]

Figure 2. Broad Usage of DevOps, Agile Software Methodologies, and Low-code/No-code Processes



Source: ESG, a division of TechTarget, Inc.

There is also a clear relationship of DevOps methodologies with delivering modern applications. ESG research shows a strong correlation between cloud-native and extensive usage of newer application development methodologies and

[5] Source: ESG Complete Survey Results: *2022 Technology Spending Intentions*, November 2021.

processes. And tying this to container adoption, ESG research shows more than one-third of organizations (38%) currently use containers for production applications.[6]

Storage Class Considerations to Support AI/ML on HCI

It would be a miss to not consider the storage class impacts to AI/ML initiatives. Storage class memory (SCM), sometimes known as persistent memory, is the latest storage impact to improve and increase performance in modern data storage systems. SCM offers a different approach to the use of NAND, extending the capacity of a standard DIMM by using NAND as storage and DRAM as a cache for active data. Both persistent memory and SCM offer high-speed approaches to data storage.

Leveraging SCM to support AI/ML workloads introduces significant benefits in response time and performance when accessing data. When compared to DRAM, SCM offers increased density, lower power consumption, and improved affordability. SCM can be used in the storage array back-end to provide a caching tier for hot data with ultra-low latency access times, a highly desirable capability to support demanding AI/ML workloads. SCM in the cloud also increases value-added performance. Distributed and multi-cloud applications leverage the performance benefits of SCM. Cloud-based data centers offering data residing on SCM tiers reduce the latency and gain the same performance characteristics found across the full storage hierarchy.

In relation to AI workloads, it is important to highlight acceleration technologies and specifically GPUs to accomplish the rapid results demanded by unique AI workloads, such as training and inference.

- Training – Fueled by massive volumes of data and requiring the use of highly performant compute, storage, and networking, training is focused on the development of algorithms to ensure high accuracy when put into production.
- Inference – Once a model is deemed acceptable, enterprises must then deploy the model to a production environment for inferencing. Inferencing has a different set of requirements than training in that it does not require heavy processing resources like training, but for many use cases, it does require fast execution of data analysis, returning a result in as close to real time as possible.

While GPU integration may sometimes be viewed as a complex integration point, success is being found by pairing GPUs with a consistent infrastructure layer to simplify deployment, optimization, automation, and utilization. A great example can be seen by incorporating GPUs into HCI platforms, where organizations can gain access to a scalable infrastructure with the horsepower to satisfy diverse AI workloads. Another similar approach is to leverage compute express link (CXL), which provides a high-speed interconnect for central processing from the CPU to the storage. This approach leverages the PCI express (PCIe) backplane for I/O, memory, and cache. This open-standard approach to increased storage access is another way to leverage high-speed access to data for AI/ML projects. It is also important to note that GPUs play an integral part in AI/ML workloads.

VMware Cloud Foundation for the Modern Enterprise

VMware Cloud Foundation with Tanzu (VCF with Tanzu) enables organizations to manage VM and container-based workloads with VMware's hybrid cloud platform, built on full stack hyperconverged infrastructure (HCI) technology. With support for both traditional and modern enterprise applications, VCF provides a complete set of highly secure software-defined services for compute, storage, networking, security, Kubernetes, and cloud management to increase enterprise agility and flexibility with consistent infrastructure and operations across private and public clouds. Paired to VMware

[6] Source: ESG Survey Results, *Supporting AI/ML Initiatives with a Modern Infrastructure Stack*, May 2021.

vSphere with Tanzu, IT and DevOps can embrace a consolidated VM and container management environment rooted in on-demand resource availability, intrinsic security and lifecycle management, live migration, and load balancing.

Delivering an AI-ready Infrastructure

In the latest release of VCF, VMware and NVIDIA are delivering an AI-ready enterprise platform. The foundation is built on VCF to help deliver a consistent architecture with simplified management. IT administrators are empowered to provision self-serve capabilities quickly and easily to their data scientists and DevOps teams when building AI/ML data pipelines that utilize vGPUs as resources within VCF. AI stakeholders can rapidly configure and deploy AI workloads with increased utilization and reduced costs by utilizing advanced features of the NVIDIA AI Enterprise Suite paired with VMware's multi-instance GPU capabilities.

Multi-instance GPUs are delivered by taking a physical GPU and partitioning it in up to seven isolated instances for optimized utilization and quality of service. This enables the sharing of GPU resources across users and workloads to improve utilization and avoid wasted GPU cycles. Utilizing VMware's Distributed Resource Scheduler (DRS), IT can ensure optimal workload placement, nondisruptive operations, optimal resource consumption, and mobility via VMware vMotion to other hosts that support the same GPU technology. Paired with persistent memory systems, the result is improved resiliency and scale of memory-intensive AI workloads and applications while eliminating downtime.

The Bigger Truth

Access to data, connectivity to data sets, and adopting new technology play key roles in the success of AI/ML initiatives. But without a foundational infrastructure stack, organizations are set up for failure in the pursuit of AI. IT teams need help to ensure right-sized resources are available to the right stakeholders when needed. They need help to ensure flash storage, SCM, and GPUs are in lockstep to address high-performance and low-latency requirements of different AI workload requirements. They need flexibility, automation, and self-service to ensure the effective scale of AI infrastructure as organizations continue to transform the business by applying AI to different use cases.

With VMware and HCI, organizations gain access to an AI-ready infrastructure building block that improves operational efficiency, easily scales, and improves AI time to value. Through access to virtualized GPU instances and container management constructs, organizations are set up for success in enabling IT to deliver an effective infrastructure that supports the dynamic concurrency demands of AI. Regardless of where your organization is on its AI journey, considering an AI-ready infrastructure to effectively scale the use of AI with your business should be on your short list.

All product names, logos, brands, and trademarks are the property of their respective owners. Information contained in this publication has been obtained by sources TechTarget, Inc. considers to be reliable but is not warranted by TechTarget, Inc. This publication may contain opinions of TechTarget, Inc., which are subject to change. This publication may include forecasts, projections, and other predictive statements that represent TechTarget, Inc.'s assumptions and expectations in light of currently available information. These forecasts are based on industry trends and involve variables and uncertainties. Consequently, TechTarget, Inc. makes no warranty as to the accuracy of specific forecasts, projections or predictive statements contained herein.

ESG a division of TechTarget

Enterprise Strategy Group is an integrated technology analysis, research, and strategy firm that provides market intelligence, actionable insight, and go-to-market content services to the global IT community.

www.esg-global.com ✉ contact@esg-global.com ☎ 508.482.0188