White Paper

Scaling Artificial Intelligence and Machine Learning Workloads

Sponsored by: VMware

Ashish Nadkarni Peter Rutten
July 2021

IDC OPINION

The business opportunities that can be achieved by investing in artificial intelligence (AI) technologies are exceptionally promising and potentially equally rewarding. Businesses know that not acting on AI is a risk that could pose an existential threat, allowing their competitors to possibly gain an edge over the business using previously unavailable data and capabilities to grow and delight their customer base. AI is here and now. Serious AI initiatives are being undertaken worldwide, across both industries and company sizes.

Until recently, many of these initiatives have been in preproduction stages. Many organizations' lines of business (LOBs), IT staff, data scientists, and developers have been working to learn about AI, understand the use cases, define an AI strategy for their business, launch initial AI initiatives, and develop and test the resulting AI applications that deliver new insights and capabilities using machine learning (ML) algorithms, especially deep learning (DL). Businesses are now ready to scale these initiatives and make AI/ML workloads a part of their next-generation workloads. IDC distinguishes three types of AI workloads:

- AI platforms (where most of the AI model training takes place)
- AI applications (the applications that are built with the completed AI model and that inference on the model)
- AI-enabled applications (applications that are partially enriched with the completed AI model and inference on it as part of their larger functionality)

IT organizations are finding out (in many cases, the hard way) that AI workloads are difficult to support. Not least of the causes is that they do not scale well on standard, multipurpose infrastructure. This is especially true not only for the deep learning training stage on AI platforms but also, increasingly, for AI inferencing when executing the model as part of an application. AI platforms parse exponentially greater amounts of data, are extremely memory bandwidth demanding, and require powerful parallel processing capabilities based on high core density — standard central processing units (CPUs) cannot execute these AI tasks. AI applications increasingly require additional core from acceleration to execute the model and often need to execute their algorithms in near real time.

Developers and data scientists can spin up instances in the public cloud to gain access to a variety of performance-intensive computing instances and try to connect them back to on-premises data repositories, but the latency between public clouds and data repositories elsewhere, the cost of graphics processing unit (GPU) or otherwise accelerated instances, and the resulting hard-to-manage silos can be detrimental to the organization's AI efficiency.

The solution is to extend or implement a hybrid cloud infrastructure for AI workloads — one that can transparently bridge on- and off-premises deployments, provide consistent access to performance-intensive hardware, and deliver seamless data access and management to these workloads. Businesses must ensure that these workloads — which are cloud native in nature — can be run in bare metal, virtualized, and containerized environments and that they can scale while providing a tangible return on investment.

SITUATION OVERVIEW

AI/ML Workloads Are Gaining Mainstream Adoption

Artificial intelligence and machine learning are achieving mainstream adoption in the enterprise. Businesses around the world are responding vigorously to the new opportunities offered by AI workloads. AI/ML capabilities provide competitive advantage to enterprises through new business models and digitally enabled products and services. They enable businesses to improve user experience, increase productivity, and innovate for the future. Throughout 2020, IDC's studies of tens of thousands of global organizations highlighted executives' shift in commitment to enterprise intelligence:

- 70% of CEOs articulated the need for their organizations to be more data driven.
- 87% of CXOs said that being a more intelligent enterprise is their top priority for the next five years.

IDC defines AI as a set of technologies that uses natural language processing (NLP), image/video analytics, machine learning, knowledge graphs, and other technologies to answer questions, discover insights, and provide recommendations. These systems hypothesize and formulate possible answers based on available evidence, can be trained through the ingestion of vast amounts of content, and adapt and learn from their mistakes and failures through retraining or human supervision.

Machine learning is a subset of AI techniques that enables computer systems to learn and improve their behavior for a given task without having to be programmed by a human. Machine learning models are algorithms that can improve over time by testing themselves repeatedly using large amounts of structured and/or unstructured data until they are deemed to have "learned" a task (e.g., recognizing a human face). Deep learning (DL) is a subset of ML, and typical DL architectures are deep neural networks (DNNs), convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), and many more.

Broadly speaking, AI/ML workloads are primarily deployed for achieving one or more of three key business outcomes:

- **Improve the end-user or customer experience of existing products or services.** This could be enhancing existing features or functions or delivering a new set of capabilities altogether. Examples include enhancement of existing services with automated and interactive digital assistants, capabilities to automatically identify objects and recommend end-user actions, and use of AI/ML algorithms in autonomous vehicles to transform mobility for humans and freight.
- **Improve operations within the organization.** This could be to improve people, process, and technological efficiency for any internal or external business or IT operations. It could be also used to ensure proper governance and compliance requirements are met. In some industries, this could mean preventive measures to avoid or minimize the risk of unplanned downtime. Examples include preventive maintenance for improving operations resiliency and inventory management for ecommerce websites.

- **Improve customer support and engagement.** This has applicability across a wide spectrum of industries. For example, in the banking industry, AI-based learning programs understand customer needs and problems that help reduce the time and resources required to resolve customer issues. AI-based customer relationship management (CRM) systems can understand customer context in real time and recommend relevant actions to sales agents. Examples include the use of chatbots on product or services websites and voice recognition–based automated phone agents to improve customer support.

Crucially, AI/ML algorithms are applied horizontally across these three functions. In many organizations, this involves data scientists or developers working to develop new custom cloud-native AI/ML applications that sit adjacent to or integrate with existing and current-generation applications and large data repositories, which are often located on premises.

During the AI model training stage, businesses are compelled to find performant infrastructure for larger and more complex AI models. IDC has seen enterprise infrastructure evolve from standalone accelerated bare metal servers to increasingly high-performance computing (HPC)–like tightly connected server clusters for such AI training purposes. Some of the most popular AI models, those for natural language processing, may consist of tens of billions of parameters, which means that, just as with modeling and simulation, performance is key. Smaller AI models still require the kind of performance that allows data scientists to iterate as often as needed without wasting time waiting for training runs to complete.

Many businesses have reached the next stage of AI development where they are deploying AI at production scale, with the AI model being inferenced on in close functional coordination with other enterprise applications. This scenario also requires careful consideration and selection of the various infrastructure options, which are not necessarily the same as those for training the model. Lighter and virtualized accelerators may well carry the inferencing workload more than sufficiently. Most important in both cases — AI training and inferencing infrastructure — is that the organizational AI/ML initiatives do not become a costly endeavor with disparity between the investment and the return on that investment.

Three Dimensions of AI/ML Workloads: Scale, Portability, and Time

AI/ML workloads could be applications that are custom developed by an organization, may be based on commercial AI software, or may be delivered as AI SaaS. Deployment considerations for the custom-developed and commercial software are on premises, in the cloud on IaaS, or as a hybrid cloud, wherein the on-premises environment interacts with a public cloud environment (e.g., apps in the cloud accessing databases on premises).

Further, for the various deployment scenarios, solutions must be considered for:

- **Training** is necessary for securely processing the volume of data that is required for training AI models with extremely high performance. The performance requirements for deep learning training involve the ability to execute massively parallel processing using GPUs or other special-purpose processors dedicated to AI training, combined with high-bandwidth data ingestion.
- **Inferencing** is required for securely processing the volume of data that the AI model will perform inferencing on with extremely high performance. Performance with respect to inferencing means the ability to process incoming data through the trained AI model and deliver near-real-time AI insights or decisions.

The key infrastructure requirements for *AI training workloads* are:

- **Performance.** Performance of the infrastructure means that a training run can be completed within a reasonable time frame from a data scientist perspective. The faster that a training run can be completed, the more iterations are feasible to improve the model's quality. It also dictates how large and complex the AI model can be, which translates directly into the sophistication or end-user value of the AI application that will be developed from the model.
- **Scale.** The scale dimension describes the scale at which the workload operates. The foundational subdimensions — compute, networking, and data persistence (storage) — are all hardware related. Software-related subdimensions such as orchestration are being introduced for maintaining balance with an increase in the size and complexity of the stack.
- **Cost.** Training AI models is an iterative process that must be financed with an opex or a capex model, or a combination thereof. When organizations experiment with, train, and iterate with AI models in the cloud, costs can quickly escalate. When they do so on premises, hardware acquisition costs can accumulate.

The key infrastructure requirements for *AI inferencing workloads* are:

- **Scale adjacent to enterprise applications.** The foundational subdimensions — compute, networking, and data persistence (storage) — are hardware related, but the demands that AI inferencing workloads place on the infrastructure are more distributed and less data intensive than those of AI training workloads. Software-related subdimensions such as orchestration are more common in the case of inferencing for maintaining balance with an increase in the size and complexity of the stack.
- **Portability.** This is the ability of the workload to be moved across core datacenter, cloud, and edge deployments. Today, many of these workloads are static in nature (i.e., designed to run in a single deployment). Increasingly, companies are looking at developing workloads in one deployment (e.g., public cloud) and installing them (in production) at another one (e.g., edge). This is analogous to the current model of developing and deploying mobile apps.
- **Time.** This relates to the time continuity of the workload itself. Many AI workloads borrow their design from high-performance computing or big data and analytics deployment — they are designed to be batch in nature. Increasingly — and thanks to the proliferation of high-performance accelerators — AI inferencing workloads can analyze streaming data in a real-time or a near-real-time manner.

AI/ML Workloads Rely on Infrastructure for Optimized Outcomes

AI/ML workloads are extremely complex not just because of the algorithms they use but because of the way they depend on infrastructure to deliver time to value. AI/ML workloads have numerous and stringent requirements. First, AI needs a lot of powerful hardware resources such as GPUs, flash memory, and high-bandwidth network. Second, enterprises need the right platforms, frameworks, data sets, and even pretrained AI models. Any imbalance in this infrastructure can have one of two consequences: It can lead to either suboptimal outcomes resulting in a failed initiative or, worse still, wasted expenses on infrastructure. Neither outcome bodes well for the business.

It is therefore important for IT organizations to be brought into the conversation early and to be an integral part of the conversation around the infrastructure requirements for AI workloads. These requirements could be summarized as follows:

- **Compute.** Most AI/ML workloads are cloud native in nature and are deployed in containers. However, many AI workloads can also be hosted on bare metal or virtualized compute instances. Many AI workloads are optimized to utilize accelerators, especially in situations where the algorithms benefit from massive parallelization. However, investments in accelerated compute can quickly add up; therefore, the selection of the proper type of accelerators is important. Crucially, this accelerated compute must be accessible to all AI/ML workloads in a shared manner, like how virtualization and containerization deliver shared compute and memory resources.
- **Data persistence.** If the compute requirements for AI workloads vary, so do the data persistence requirements. An underrepresented and misunderstood aspect of the AI workload stack is the data persistence layer. It is often assumed that all AI workloads require a large amount of high-performance storage. The fact of the matter is that not all AI workloads are "big data sets" — they may be sampling lots of small data sets concurrently for a short period of time.
- **Deployment locations.** Given the distributed nature of AI/ML workloads, it is much simpler for the algorithms to shift closer to where data is being sourced or generated than the other way around. In recent times, the cloud-core-edge approach has become a de facto model for AI/ML deployments wherein diverse workload profiles can be matched to the location in a seamless manner while minimizing data movement. In other words, make it easier for compute to move closer to the data and not the other way around.

The AI/ML Hybrid Cloud Stack

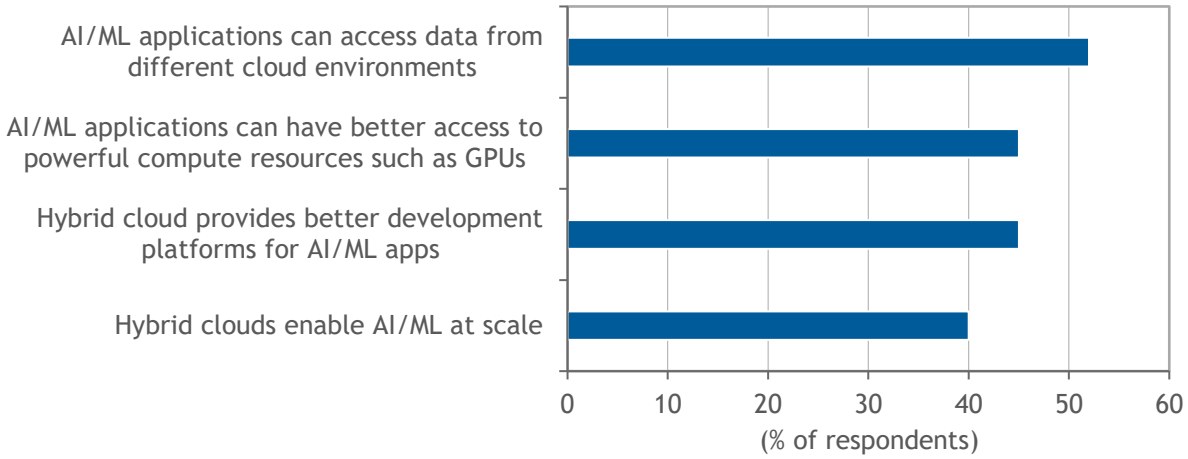
IDC believes that the starting point for any AI/ML workload production (i.e., inferencing) deployment must be a hybrid cloud infrastructure strategy, which is implemented using best-of-breed hardware and software platforms. Hybrid cloud infrastructure is becoming the gold standard for IT architecture and refers to an environment where enterprise customers have a single control pane to monitor and manage multiple private clouds, public clouds, and even legacy infrastructure. In such an environment, applications can seamlessly migrate from one cloud to another, and data residing in any cloud can be accessed from any other cloud in a frictionless way.

Figure 1 illustrates the key benefits of deploying AI/ML workloads in a hybrid cloud infrastructure. A hybrid cloud offers an ideal environment with higher scalability and cost optimization for deploying and scaling AI/ML workloads. Thus AI models and applications can access the data from different sources without having to engage in time-consuming and expensive migration of data. Further, a hybrid cloud enables operationalizing AI at scale, especially when AI applications are built with cloud-native technology.

FIGURE 1

Hybrid Cloud Infrastructure Accelerates AI/ML Production

Q. What unique benefits do hybrid clouds offer for your AI/ML projects? Select all that apply.



n = 1,328

Source: IDC's *Cloud Pulse,* 2Q20

In a hybrid cloud environment, infrastructure platforms and systems are stacked as a modular set of layers that interoperate seamlessly. A hybrid cloud environment is inherently software defined and infrastructure agnostic. Modularity and interoperability are the key pillars of any hybrid cloud solution — they enable the IT organization to embrace a vendor-agnostic approach to creating their own version of hybrid cloud using products and services from leading vendors. The three essential layers of a hybrid cloud optimized for AI/ML workloads are workload layer, computing platform (processor) layer, and software-defined infrastructure (SDI) layer.

Workload Layer

- The workload layer is where businesses either modernize existing applications or build new applications. There are three ways that businesses embrace AI/ML in their environment:
  - The AI software and platforms category of AI workloads refers to the tools and platforms used to build or implement AI capabilities. Essentially, these are AI training platforms.
  - The AI applications category refers to workloads that are used to deliver AI capabilities and in which AI technologies/algorithms are central to their functionality. In other words, the application cannot function as intended without AI capabilities. These are AI inferencing applications.

- The AI-enabled applications category refers to workloads enhanced and enabled by AI capabilities. In an AI-enabled application, the application vendor enhances the functionality of an existing application by introducing new AI capabilities without altering the core functionality or behavior. If AI technologies were removed from an AI-enabled application, they would still be able to function, albeit less effectively. These are applications with some amount of AI inferencing.

There are plenty of tools to build and deploy these workloads. They include:

- IDE workflow makes it easier to develop models/code.
- Application deployment framework (ADF) provides a complete set of tools and libraries to build AI applications, including any libraries, SDKs, or reference code.
- Application optimizer tools such as Kubeflow help deploy, scale, and manage ML models easily. They align closely with the ML data life-cycle management pipeline.
- Model libraries provide a curated marketplace type of experience to share prebuilt ML models.

Note: IDC refers to a "workload" as a set of applications along with their primary and secondary data sets. Workloads enable specific business outcomes, such as collaboration or content management.

Computing Platform (Processor) Layer

The foundation for any infrastructure is the hardware layer, and it is no different for AI/ML workloads. However, the operative word for infrastructure used for AI/ML workloads is "heterogeneous." Having spent much of the past decade standardizing on homogeneous infrastructure, IT organizations must now contend with the increased diversity of processors used in infrastructure for AI/ML workloads:

- **General-purpose compute — processors and memory.** Performance-optimized processors used for AI/ML workloads could include high core count x86 variants from vendors such as Intel and AMD, each with its own requirements and support for memory, I/O, and data persistence (e.g., NVMe Flash) and networking. For many AI/ML inferencing workloads, general-purpose compute resources can provide an economical alternative to accelerated hardware, but increasingly, AI/ML workloads are relying on special-purpose compute, certainly for AI training.
- **Special-purpose compute — accelerated computing.** Accelerated computing resources include graphic processing units, field-programmable gate arrays (FPGAs), and application-specific integrated circuits (ASICs). GPUs, though originally not designed to support AI/ML workloads, have been particularly suitable to run AI model training as part of an AI pipeline owing to their highly parallel structure. GPUs are easy to program and provide excellent performance but can sometimes be expensive. Lighter GPUs have been developed to facilitate AI inferencing. FPGAs are used mostly for inferencing at scale, and multiple ASICs are being developed to support either AI training or AI inferencing. On the other hand, GPUs are easy to program but expensive, FPGAs are slightly harder to program and generally somewhat less expensive, and ASICs are hard to program, requiring extensive and costly development time, but are cheap on a per-unit basis.
- **Special-purpose compute — function offload accelerators (aka Smart NICs).** These new-generation devices allow for certain kernel-level functions such as networking, security and, in some cases, even data persistence to be offloaded to a dedicated layer. By converting hardware calls to software-defined calls, these devices provide a secure, distributed computing (disaggregated hardware) layer. This disaggregated hardware layer can then be composed by hypervisors to support bare metal, virtual machines, and container instances.

Software-Defined Infrastructure Layer

The SDI layer acts as the bridge or interconnect between the workload and hardware layers. SDI brings a variety of benefits to the stack, including the ability to scale the infrastructure across on and off premises in a transparent manner. The SDI layer includes:

- **Virtualization** enables IT organizations to deliver a consistent, scalable, and shared experience to all AI production workloads in their datacenter and in the public cloud. Virtualization enables IT organizations to extend the consistency of service experience and quality to accelerated computing. This is important because accelerated computing can quickly become an unwieldy experience given the plethora of accelerators, each with their customized software stacks. Placing all this translation and optimization within the SDI layer enables the sharing of pooled accelerated computing resources across multiple AI/ML workloads and improves maintenance and upgrades.
- **Data access and management** provides a consistent (and integrated) data layer to the AI/ML workloads, enabling them to access data across deployments using file, block, object, and streams protocols. For some models, high-performance flash access is necessary. Data services include data mobility, data protection, copy data management, and tiering based on data decay. Model data life-cycle management ensures the proper treatment of data used by AI/ML models as it goes through five phases — ingestion, sanitization, train, test, and infer. Data ingestion refers to collecting the data that is used by the models. Sanitization refers to cleaning up the data to ensure good quality of data or the removal of any sensitive information. Data management is especially important in a cloud-core-edge model where the workloads move to the data. That means that data services need to operate in a workload-aware manner.
- **Orchestration and automation software** includes a variety of software that enables cross-platform and cross-premises deployments. Some examples are:
  - **Hardware resource management.** This enables the management of compute, storage, and networking hardware resources and the allocation of these resources to the virtualization layer. This is usually part of the hypervisor management console. An integrated data persistence layer enables access to persistent storage through the constructs of logical volumes and persistent volumes.
  - **Containers and orchestration.** Container orchestration platforms such as Kubernetes provide a scalable, consistent, and interoperable environment to run, deploy, and manage containers. These platforms also provide the ability to build a cluster of resources underneath and, therefore, integrate with the hardware resource management layer.
  - **Accelerated compute optimization layer.** This provides platforms, tools, and abstractions to leverage accelerated hardware. For example, GPUs are highly parallel in structure and are well suited for massively parallel applications. They need parallel programming platforms and models to be able to run these parallel applications.

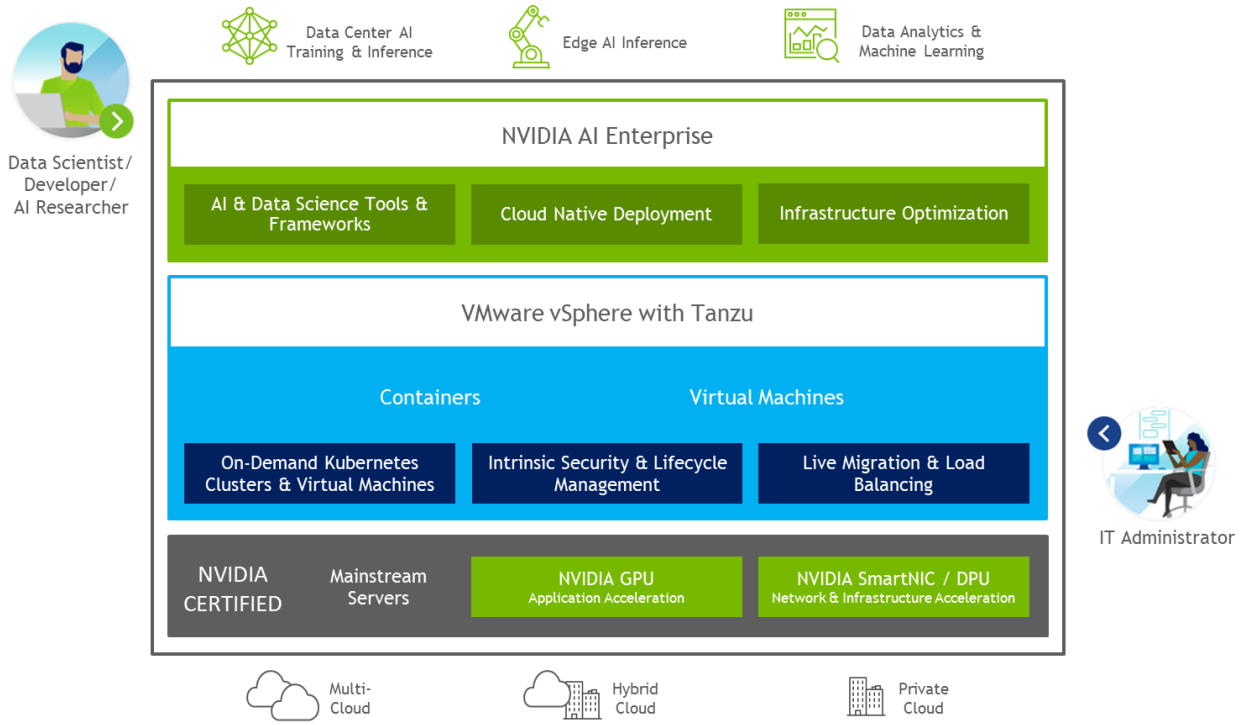
VMWARE HYBRID CLOUD SOLUTIONS FOR AI/ML WORKLOADS

VMware – a leading vendor in hybrid cloud solutions – is making it simpler for IT organizations to deploy AI/ML workloads in much the same way they would deploy other enterprise or business workloads. Figure 2 illustrates the following key aspects of the VMware solution:

- From the business side, developers and application owners benefit from a more streamlined experience. They gain the ability to run these workloads on purpose-built and optimized infrastructure, complete with NVIDIA Enterprise AI software and high-performance accelerated compute, high-performance flash storage, and as part of a hybrid cloud environment for cloud-native AI workloads.
- From the IT side, the operations team can manage the AI/ML environment in the same manner as the rest of the IT infrastructure, using the same tools and processes that it would use for provisioning and management. IT organizations benefit from not creating another silo and not requiring any special skills to manage it.

FIGURE 2

VMware AI-Ready Enterprise Platform



Source: VMware and NVIDIA, 2021

Accordingly, VMware is taking a full-stack approach that integrates key aspects of the AI/ML infrastructure stack with its existing hybrid cloud infrastructure solution called VMware Cloud Foundation. Since the support and optimization for the accelerated hardware are built into the VMware vSphere kernel, there is nothing special that IT organizations need to do when the workloads are moved from one deployment to another. The same applies to data management, which is delivered via the VMware vSAN layer that is part

of the stack. Data scientists and developers can now easily create and test AI/ML workloads in a public cloud service such as Amazon Web Services, Microsoft Azure, or Google Cloud Platform and then move them back on premises using the same set of VMware tools they would use for their other workloads. VMware continues to expand its support for various accelerators and their associated software libraries. It is also expanding other aspects of the stack including the ability to deploy bare metal as well as virtualized and containerized workloads in a hyperconverged or composable manner on disaggregated hardware. VMware is also making it easier to support AI/ML on premises with an NVIDIA partnership that was announced in September 2020 and that was established to optimize NVIDIA AI Enterprise software and GPUs for VMware's virtualization layer.

VMware vSphere 7 Update 2 (vSphere 7U2)

Released in April 2020, VMware vSphere 7 was re-architected into an open container-native platform using Kubernetes to provide a cloudlike experience for developers and operators. With more recent updates to vSphere 7, additional important milestones have been achieved, especially for running AI/ML workloads. These include:

- **Support for the NVIDIA AI Enterprise software suite, an end-to-end, cloud-native, suite of AI tools and** frameworks to streamline development and deployment of AI workloads on VMware vSphere. NVIDIA's Ampere generation of GPUs was added, including the A100, which, according to NVIDIA, comes with up to 20x better performance versus the previous generation. vSphere 7 also supports NVIDIA's multi-instance GPU (MIG) partitioning technology directly within the vSphere layer, so workloads can share A100 GPUs and benefit from higher GPU utilization, better data scientist access to available GPUs, security benefits from partitioning, fault isolation, easy migration, and load balancing.
- **Improved data security.** vSphere 7 improves data security, which is an important consideration for AI/ML workloads that operate on sensitive data sets. It also includes an improved built-in key management system and delivers confidential containers through support for on-chip encryption from AMD for its EPYC line of processors.
- **Improved data access and life-cycle management.** VMware has extended its vSAN layer to make it simpler for all data access and management to be consolidated and controlled via a single control plane. VMware has also extended the vSphere Lifecycle Manager to support vSphere with Tanzu.

VMware vSphere 7 has evolved to incorporate new capabilities to address the needs of businesses, making VMware Cloud Foundation a compelling hybrid cloud platform for deploying AI/ML workloads at the core (on premises), in the public cloud (via VMware's partnerships with AWS, Google, IBM, and Microsoft), and at the edge.

VMware Project Monterey

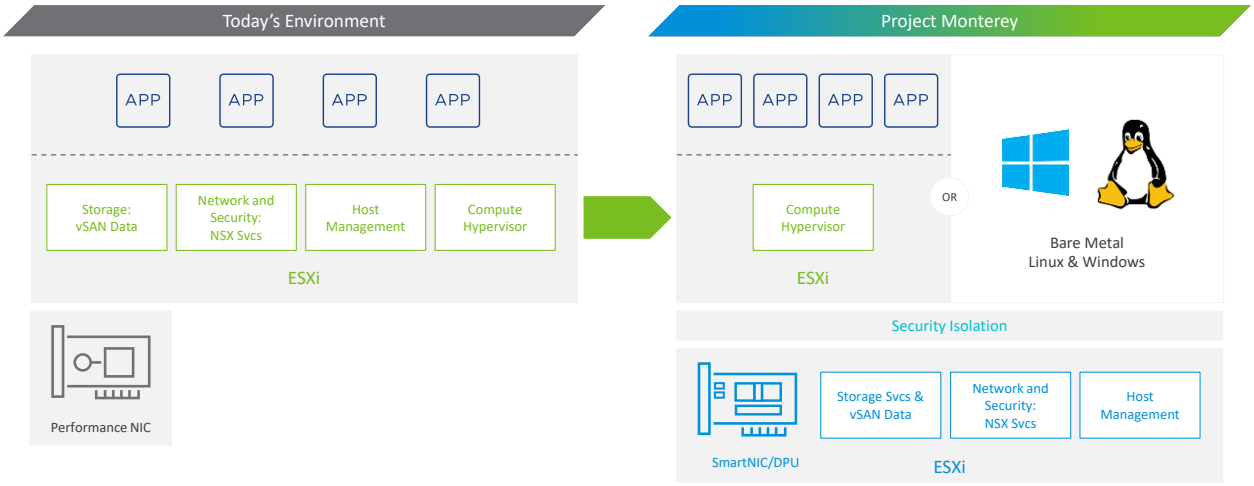
VMware's Project Monterey announced as a technology preview during VMworld 2020 is an evolutionary architectural approach for the datacenter, cloud, and edge to address the changing requirements of AI/ML workloads. As Figure 3 illustrates, Project Monterey modifies the way VMware Cloud Foundation (VMware vSphere, VMware vSAN, and VMware NSX) runs inside a server to make use of additional coprocessor (function accelerator) cards placed inside an existing server. The initiative consists of three key elements:

- **Support for function accelerator cards (FACs) from an ecosystem of partners.** VMware Cloud Foundation will be able to maintain compute virtualization on the server CPU while offloading networking and storage functions to a specially designed coprocessor on the FAC. This will allow applications to maximize the use of the available network bandwidth while saving server CPU cycles for top application performance.

- **Platform re-architecture.** VMware will re-architect VMware Cloud Foundation to enable disaggregation of the server including extending support for bare metal servers. This will enable an application running on one physical server to consume hardware accelerator resources such as GPUs or FPGAs from other physical servers. This will also enable physical resources to be dynamically accessed based on policy or via software API, tailored to the needs of the application. In addition, organizations will be able to use a single management framework to manage all their compute infrastructure, whether virtualized or bare metal. The decoupling of networking, storage, and security functions from the main server allows these functions to be patched and upgraded independently from the server.
- **Security.** With Project Monterey, VMware can implement intrinsic security. Each FAC can run a full-featured stateful firewall and advanced security suite. Since this will run on a card and not on the CPU, up to thousands of tiny firewalls will be able to be deployed and automatically tuned to protect specific application services that make up the application. Wrapping each service with intelligent defenses can shield any vulnerability of that specific service. This will enable a custom-built defense that can be automatically tuned and deployed across tens of thousands of application services. In addition, Project Monterey will enable enterprises or service providers supporting multiple tenants to isolate them from the core infrastructure.

FIGURE 3

VMware's Project Monterey



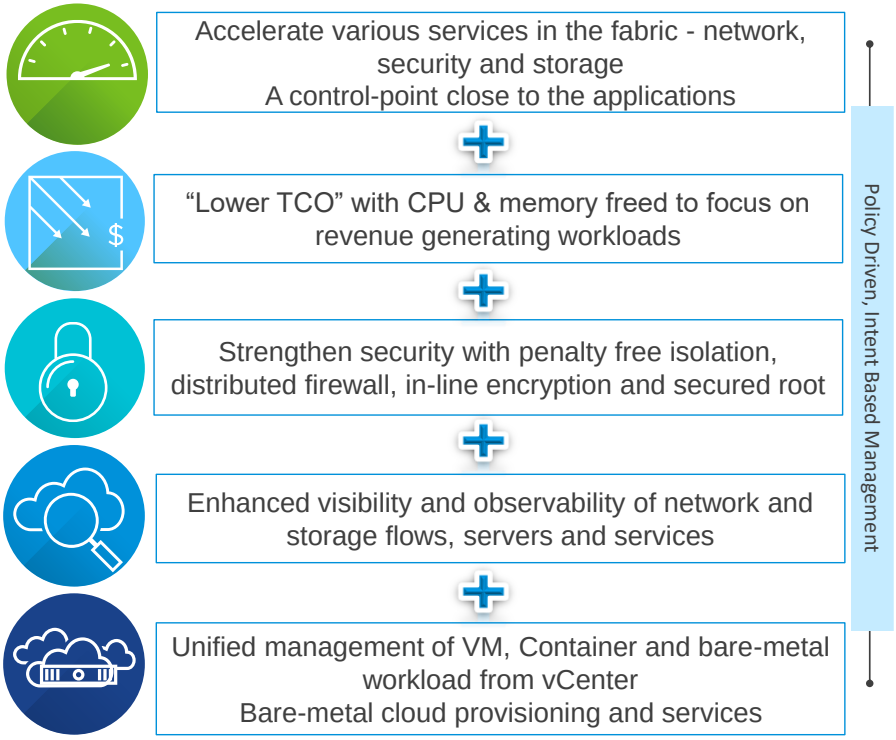
Source: VMware, 2021

VMware announced that it is taking an ecosystem approach with Project Monterey, which means that it will support FACs from several providers such as Intel, NVIDIA, and Pensando Systems inserted into or integrated with server solutions from OEM vendors such as Dell Technologies, Hewlett Packard Enterprise, and Lenovo.

Figure 4 illustrates key aspects of the VMware Project Monterey. Chief among them are consistent services, increased security, lower TCO, enhanced operator visibility, and unified management capabilities.

FIGURE 4

Benefits of VMware Project Monterey



Source: VMware, 2021

VMware and NVIDIA Address Challenges in Implementing AI/ML Infrastructure On Premises

IDC's ongoing research on AI infrastructure has uncovered many challenges that IT organizations face when embracing AI infrastructure at scale. The sections that follow discuss the ways that VMware's hybrid cloud solution and the NVIDIA AI enterprise software suite address some of these challenges. Figure 5 illustrates the VMware-NVIDIA solution, which is a complete suite for AI and data science, optimized for vSphere, and supported with regular updates for AI.

Difficult to Manage and Scale and Lack of Interoperability

Many IT organizations start by creating an AI infrastructure that is isolated from the rest of their IT infrastructure. They cite reasons such as unfamiliarity with the hardware accelerators and the capabilities of associated software stacks that require more frequent updates to the infrastructure and workload performance and scaling requirements to keep this environment separate. Unfortunately, this creates a situation where the environment itself becomes a "white elephant" that is expensive to procure and challenging to maintain.

VMware's Solution — Integrated Support for Accelerators in vSphere

VMware integrates common software optimization and middleware stacks required to interact with accelerators into the vSphere kernel. This includes the technologies that enable these accelerators to be partitioned and virtualized, so they can be shared by multiple AI/ML workloads without requiring the environment to be completely overhauled between workloads. For example, VMware has integrated NVIDIA's MIG partitioning technology for Ampere GPUs with vSphere 7. Now, IT organizations can simply extend their virtualization environment (with which they are intimately familiar) to manage their AI/ML infrastructure.

Managing Workload Mobility Between Locations and Adjacency Limitations

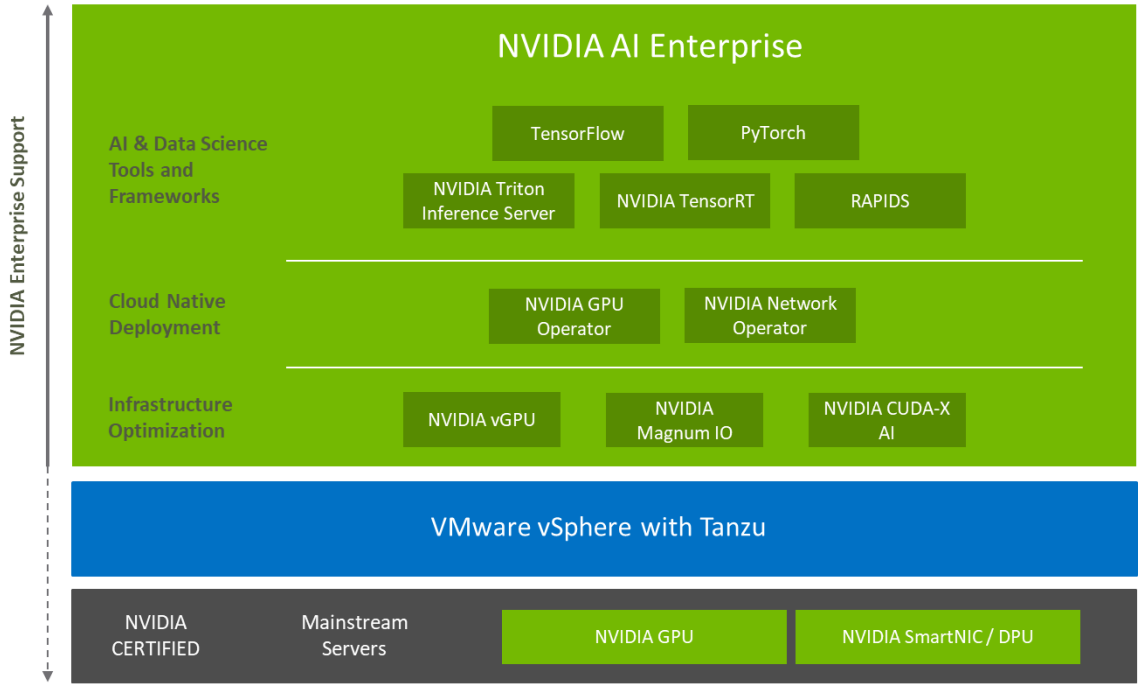
As mentioned previously in this white paper, it is not uncommon for developers and data scientists to develop and deploy AI/ML workloads in the public cloud to get around difficulties in procuring the appropriate hardware on premises. The plethora of options offered by public cloud service providers is certainly appealing. Owing to governance restrictions, however, these workloads interact with data repositories that are located on premises. Eventually, when the time comes to move these workloads into production, they must be moved back on premises, and this is where the lack of parity and consistency between environments can present a challenge to IT organizations.

VMware's Solution — Hybrid Cloud with VMware Cloud Foundation

By deploying a hybrid cloud infrastructure, IT organizations can create a uniform workload deployment and presentation layer for their constituents. Developers and data scientists can now get the best of both worlds: They can develop AI/ML workloads in the public cloud and then move them to on premises or edge locations seamlessly. They gain the ability to self-service their needs for on-demand GPU-powered Kubernetes (K8s) clusters while the tools and software stacks remain the same across deployments. Service quality attributes such as performance, scaling, and data and workload adjacency are preserved in the process given that they do not have to directly deal with the underlying hardware anymore. An additional benefit is that data access and management challenges are addressed because the stack has end-to-end support for performance- and capacity-optimized data, without having to "leave" the environment. In addition, the value of the joint platform is that AI and data science development can be supported from the core on-premises datacenter should the requirements dictate.

FIGURE 5

VMware-NVIDIA Solution for Implementing AI/ML Infrastructure On Premises



Source: VMware and NVIDIA, 2021

ESSENTIAL GUIDANCE

For IT Buyers

The time is now for organizations to envision, design, and deploy a hybrid cloud environment that delivers a consistent and scalable service quality for their workloads, including their AI/ML workloads. The selection of the proper algorithms and software development approaches is important when investing in the development of AI/ML workloads. However, the hardware- and software-defined infrastructure stack is equally important for the following reasons:

- From data preparation to model development to runtime environments to training, deploying, and managing AI models, the requirements for the underlying infrastructure defy the old models of general-purpose hardware and software platforms. Only infrastructure that is designed for data-intensive workloads with superior core performance, multiple GPUs or other forms of acceleration, fast interconnects, large amounts of coherent memory, and exceptional I/O bandwidth can execute AI/ML workloads fast enough.
- Old rigid models of running workloads in a single deployment model (at the core, in the cloud, or at the edge) do not scale well for AI/ML workloads. This is important because for data scientists and developers, it can sometimes be easier to start an AI initiative in a public cloud, and then move the developed models back on premises – or, in some cases, maintain the computing and data environments across two locations. Orchestration and automation software make it easier for data scientists and developers to stitch multiple workloads together to develop composite AI/ML workloads.

Organizations will need to make decisions about replacing or augmenting existing general-purpose infrastructure with a stack that can scale well for AI/ML workloads. A hybrid cloud infrastructure will enable businesses to develop and deploy cutting-edge AI/ML workloads.

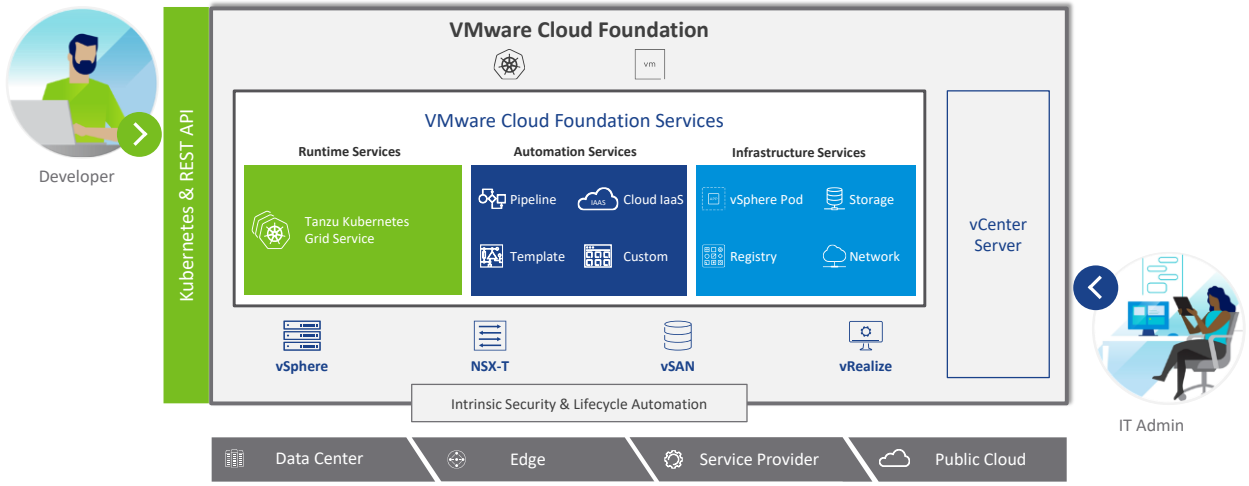
For VMware

Figure 6 illustrates how VMware is steadfastly extending its value proposition for AI/ML workloads by expanding the capabilities of VMware Cloud Foundation. In doing so, it seeks to make IT administrators the heroes in the quest of their business to differentiate itself by applying AI/ML algorithms to improving business operations and customer experience and support. VMware enables IT organizations to comprehensively bridge their on-premises and cloud-based environments for AI/ML and next-generation workloads.

Longer term, and as initiatives such as Project Monterey are folded into the core VMware Cloud Foundation solution, VMware can introduce a refined approach for implementing composable/disaggregated infrastructure at scale. For example, Project Monterey will enable organizations to adapt datacenter, cloud, or edge environments for application-specific performance, availability, and security needs. In addition, the initiative will extend VMware infrastructure and operations for all applications – reducing the need for specialized systems, teams, and management tools – which in turn will be able to reduce overall complexity and TCO for IT teams. VMware can also enlist its extensive and growing ecosystem of OEMs, accelerated compute providers, and independent infrastructure software vendors and define a de facto path forward in which entities with disparate interests can operate under a common hybrid cloud framework. It can bring solutions to the market that redefine (and ultimately blur) the way the datacenter delivers its services to end users.

FIGURE 6

VMware Cloud Foundation Hybrid Cloud Stack for AI/ML Workloads



Source: VMware, 2021

CONCLUSION

Businesses need a consistent hybrid cloud strategy to deploy and scale AI/ML workloads — workloads that they will rely on as they embark on the next phase of their digital transformation strategy.
By extending vSphere and VMware Cloud Foundation to host AI/ML workloads, VMware enables IT organizations to extend the service quality of their hybrid cloud environments to their AI/ML workloads. In other words, they can fold AI/ML workloads into their enterprise hybrid cloud, thus enabling businesses to get the most out of their investments in AI/ML workloads.

About IDC

International Data Corporation (IDC) is the premier global provider of market intelligence, advisory services, and events for the information technology, telecommunications and consumer technology markets. IDC helps IT professionals, business executives, and the investment community make fact-based decisions on technology purchases and business strategy. More than 1,100 IDC analysts provide global, regional, and local expertise on technology and industry opportunities and trends in over 110 countries worldwide. For 50 years, IDC has provided strategic insights to help our clients achieve their key business objectives. IDC is a subsidiary of IDG, the world's leading technology media, research, and events company.

Global Headquarters

140 Kendrick Street
Building B
Needham, MA 02494
USA
508.872.8200
Twitter: @IDC
blogs.idc.com
www.idc.com