White Paper

Storage Infrastructure Considerations for Artificial Intelligence Deep Learning Training Workloads in Enterprises

Sponsored by: NetApp

Eric Burgener Peter Rutten
October 2022

IDC OPINION

As enterprises move into the digital era through a process termed digital transformation, they are moving to more data-centric business models. While there are big data and analytics workloads that do not use artificial intelligence (AI), AI-driven applications are growing at a rapid rate over the next five years. AI workloads also include machine learning (ML) and deep learning (DL) workloads, and while more data helps drive better business insights across both application types, it is particularly true for DL workloads. Experience with DL workloads over the past three years indicates that outdated storage architectures can pose serious challenges in efficiently scaling large AI-driven workloads, and over 88% of enterprises purchase newer, more modernized storage infrastructure designs for those types of applications.

AI-driven applications commonly have a multistage data pipeline that includes ingest, transformation, training, inferencing, production, and archiving, with much of the data shared between stages. When enterprises can consolidate the storage processing for all of these stages onto a single storage system, it provides the most cost-effective infrastructure. This type of consolidation can, however, potentially put performance, availability, and security at risk if the underlying storage does not support low latency, high data concurrency, and the multitenant management to manage the data for each stage according to individual requirements. Successful AI workloads tend to grow very fast, making easy and high scalability important as well. Enterprises often want to use public cloud-based services during the data life cycle, so systems need to be enabled with an ability to support cloud-native capabilities and integration. And finally, because these systems tend to be quite large (often growing to multi-petabyte [PB] stage and beyond), high infrastructure efficiency is important to be able to drive a low total cost of ownership.

Because of these requirements, enterprises are gravitating more and more to software-defined, scale-out storage infrastructures that support extensive hybrid cloud integration. NetApp, a well-established storage systems vendor that held the number two market share spot by revenue for external storage in 2021, offers enterprise-class storage systems that meet these requirements well. In addition to offering the technical capabilities demanded by AI workloads, NetApp has also implemented a deployment model that makes buying IT infrastructure solutions for AI-driven workloads fast and easy.

NetApp ONTAP AI provides a prepackaged solution. The solution includes accelerated compute and networking from NVIDIA that is so often needed for AI workloads, and a software-defined, scale-out storage architecture based on ONTAP, the vendor's proven enterprise-class storage operating system,

under a single SKU. There are different single SKU configurations for specific compute, storage performance, and capacity options. With NetApp ONTAP AI, all components have been pre-validated by the vendors to work together; the system includes a unified management graphics processing unit (GPU) for simple, centralized management, and it is all covered under a maintenance contract with a single point of support contact with NVIDIA Coordinated Support. Converged infrastructure stack offerings like NetApp ONTAP AI deliver fast time to value for enterprises deploying AI workloads, and they include access to common AI tools and other components that enterprises invariably need as they harness AI to drive faster, better business decisions. Enterprises deploying AI-driven workloads would do well to consider NetApp storage as part of their AI infrastructure solution.

SITUATION OVERVIEW

AI technologies are becoming an increasingly important driver of business success for digitally transformed enterprises today. ML, a subset of AI, reviews data inputs, identifies correlations and patterns in that data, and then applies that learning to make informed decisions across a wide variety of use cases — everything from recommendation engines and fraud detection to customer analysis and forecasting events — that can supplement and help guide better human decision making. DL is an evolution of machine learning that uses more complex, multilayered neural networks that can actually learn and make intelligent decisions on their own without human involvement.

Unlike ML, DL uses a multilayered structure of algorithms whose design is inspired by the biological network of neurons in the human brain to provide a learning system that is far more capable than that of standard ML models (which are based on only one- or two-layer "network"). DL is used to address far more complex problems such as natural language processing, virtual assistants that mimic humans, and autonomous vehicle systems, and these models may have thousands to trillions of "parameters" in the multilayer neural network. IDC's *AI DL Training Infrastructure Survey,* completed in June 2022, explored existing enterprises' use of AI technology across both ML and DL workloads.

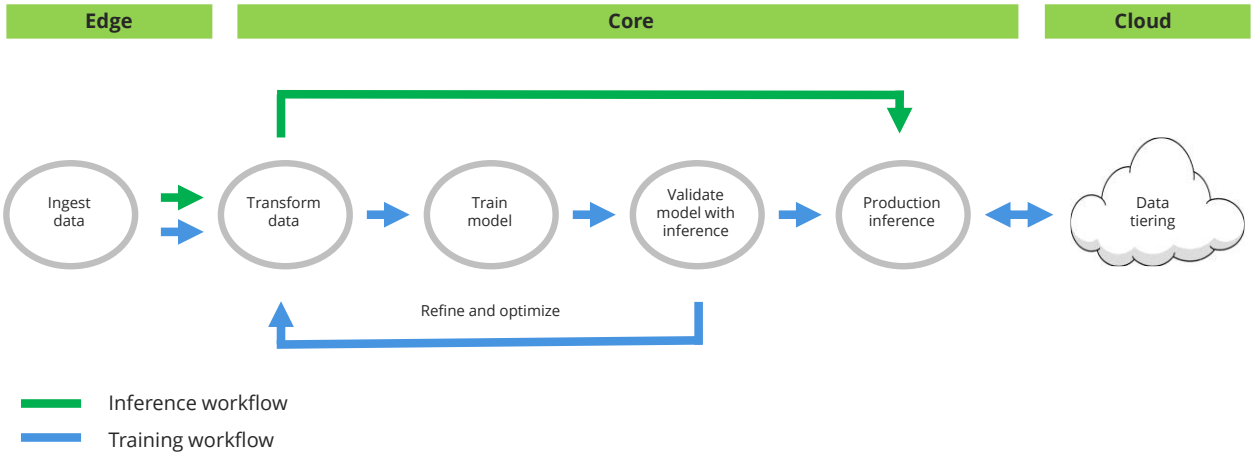
AI workloads in general perform better when they leverage larger data sets for training purposes, and this is particularly true for DL workloads. AI life-cycle applications are one of the two fastest-growing workloads over the next five years (with non-AI-driven big data analytics workloads being the other one), and they are contributing strongly to the projected data growth rates in the enterprise. Most enterprises are experiencing data growth rates from 30% to 40% per year and will soon be managing multi-petabyte storage environments (if they are not already). Roughly 70% of enterprises undergoing digital transformation will be modernizing their storage infrastructure over the next two years to deal with the performance, availability, scalability, and security requirements for the new workloads they are deploying in an era of heightened privacy concerns and rampant cyberattacks.

Figure 1 shows a typical AI data pipeline with its multiple stages. In a typical AI project, the enterprise initially carefully identifies a problem and the data science team chooses the algorithm they will use that maximizes the possibility of success. Data is then gathered and ingested into the storage infrastructure where it will be converted into a usable format and then used to train the AI model. As the model is being built through this "training" effort that runs the data set through the algorithm, the data science team is evaluating the accuracy of the recommendations or predictions the model is making (the "inferencing" stage). Once the accuracy is acceptable, the model can be deployed in production and the team will develop applications that use the model. But AI applications are not static, and the training workflow must continue to be refined and optimized over time or the usefulness of the

model will degrade (hence the need for the iterative training and inferencing workflows shown in Figure 1). Older data may or may not be archived, depending on its usefulness to the evolution of the model.

FIGURE 1

AI Data Pipeline Stages



Source: IDC, 2022

The different stages of the AI data pipeline require different capabilities from the IT infrastructure. Accelerated compute, implemented through graphics processing units, application-specific integrated circuits (ASICs), or floating-point gate arrays (FPGAs), can be much more efficient than general-purpose processors when dealing with data-intensive workloads. IDC's *AI DL Training Infrastructure Survey* indicated that 62% of enterprises running AI workloads were running them on "high-density clusters," defined as scale-out IT infrastructure leveraging some form of accelerated compute. Acceleration is based on parallelization of the data and/or the model, allowing for simultaneous processing on thousands of cores and potentially hundreds of compute systems rather than sequential processing. Parallel processing architectures are optimal for deep learning neural networks.

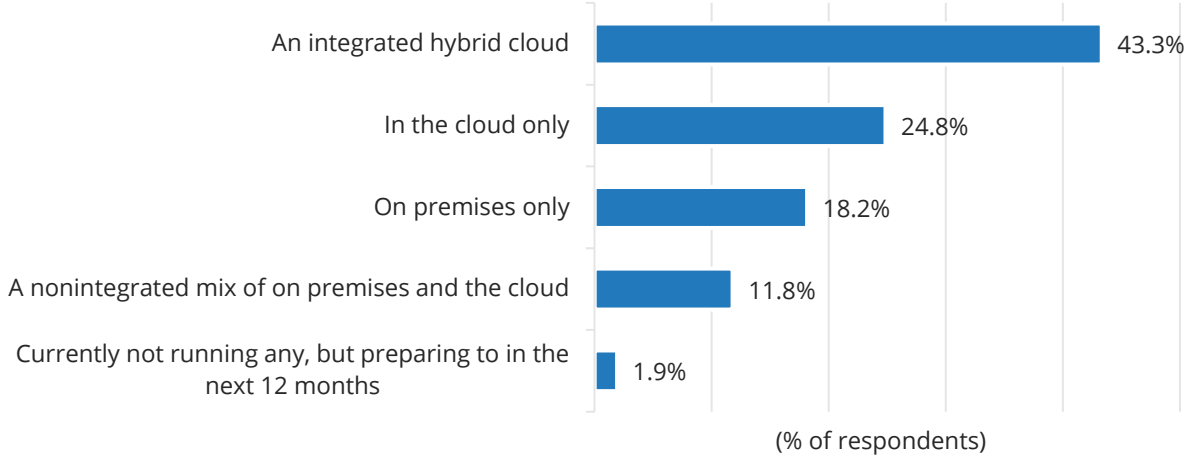
Implementing the right storage infrastructure can also make a big difference, particularly for larger-scale AI workloads. Already, over 43% of enterprises running these workloads operate in integrated hybrid cloud environments, and 69% of them are regularly moving data between on- and off-premises locations. Software-defined storage is important in providing the flexibility and data mobility needed in these hybrid cloud environments. Given the high data growth rates, scale-out architectures provide the widest range of scalability needed with easy nondisruptive expansion over time.

Figure 2 shows AI DL training projects being run in the most common deployment scenarios.

FIGURE 2

Integrated Hybrid Cloud — The Most Common Deployment Scenario

Q. Are you currently running your AI deep learning training projects in any of the following deployment scenarios?



n = 314

Base = all respondents

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

The use of managed services for AI training workloads is popular, with 62% of the enterprises surveyed using them although deployment locations are split. Over 46% of them rely on infrastructure at a managed service provider location, almost 30% of them use this approach with on-premises infrastructure, and over 8% are working with colocation partners like Equinix. Top factors driving the interest in the use of managed services include a better perceived value proposition versus more traditional "do it yourself" methods, protection against obsolescence in a rapidly changing market, and the ability to offload infrastructure management so that staff can focus on more strategic operations.

AI Training Compute Infrastructure Considerations

Compute infrastructure for AI has evolved significantly over the past half decade and today represents nearly 17% of all the servers sold worldwide per year in terms of sales. IDC expects the portion of servers that is used for AI to grow to 22% of the total server market by 2026. Increasingly, these are not general-purpose servers anymore. Rather, they are designed with higher-performance host processors; high-end co-processors such as GPUs, FPGAs, and ASICs; fast interconnects; high-speed networking; liquid cooling; and rich software stacks for AI.

As such, AI infrastructure has begun to resemble high-performance computing (HPC) infrastructure and large-scale AI systems are starting to look more and more like supercomputers, while supercomputers are increasingly used for executing AI workloads. AI and HPC are not only converging on the same infrastructure but also converging as workloads, with AI front-ending an HPC simulation to cut back on the number of simulation runs or with an AI model operating as a side process inside an HPC loop. IDC refers to these types of applications as performance-intensive computing (PIC), an umbrella term for AI, HPC, and big data and analytics workloads. PIC has upended the homogeneous, general-purpose datacenter and led a revolution of purpose-built designs not just from the traditional

server vendors but also from NVIDIA as well as from various start-ups such as SambaNova, Groq, Graphcore, and Cerebras.

The major driver for these developments is the AI model size. To achieve high levels of accuracy, it is not enough to feed an AI training model large amounts of data; the model also needs to have a large number of layers and parameters, which are the weights of the connections in the neural network. Just two years ago, the AI community marveled at model sizes that had reached 300 billion parameters; today AI scientists have developed trillion-plus parameter models.

Unfortunately, the number of parameters in a neural network and the volumes of data fed into it correlate directly with the amount of compute required. In other words, the more capable and/or accurate an AI model needs to be, the more compute is required to train that model. What's more, the faster an organization wants to have its model trained, the more compute is required as well since model training can be distributed across many nodes in a cluster; hence a larger cluster trains a model with greater speed. This is why companies like NVIDIA today build and install DGX SuperPODs in customer datacenters.

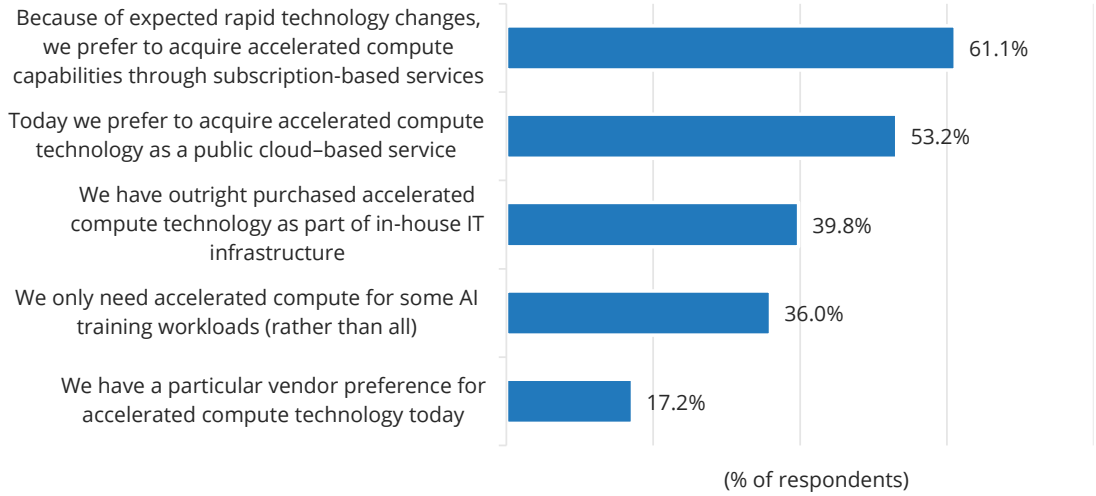
It is also the reason for two other trends: the increasing interest in foundational or pretrained models, whereby an organization does not attempt to train a complex model "from scratch" but instead purchases it and then fine-tunes it for its own purposes; and the growing interest in as-a-service AI platforms, which remove all infrastructure management from the customer and provide an AI training environment that is operated as a service and paid for with a subscription model, either on premises or in a cloud. (Note that public cloud puts significant pressure on AI scientists and engineers as every iteration of the model training adds to the cloud bill.)

Figure 3 shows enterprises' choices in regard to accelerated compute services.

FIGURE 3

Enterprise Preference for Subscription-Based Accelerated Compute Services

Q. With reference to accelerated compute (i.e., GPU-based servers) technology, please select all statements below that apply.



n = 314

Base = all respondents

Data is managed by IDC's Quantitative Research Group.

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

It is this context that should inform an organization's AI compute infrastructure considerations. Anticipated model sizes and accuracy requirements drive compute requirements. In finance or healthcare, for example, accuracy can be paramount. Some types of models are more demanding than others. Natural language processing and recommendation engines, for example, are especially compute intensive. In summary, the infrastructure choices that organizations make for their AI workloads need to be informed by the anticipated model type, number of neural network layers, the number of parameters, the anticipated data volumes, the desired speed of training completion, and the number of models to be trained simultaneously. All of this needs to then be fitted into a cost and SLA mandate.

AI Training Storage Infrastructure Considerations

Enterprises by and large realize that AI training workloads have requirements that are not necessarily well met by legacy storage infrastructure, and over 88% of them purchase a storage system specifically for the new workload. It is interesting to note, however, that 60% of enterprises use that new storage system to host other AI workloads as well, potentially raising issues around performance (i.e., noisy neighbors), an ability to support high degrees of concurrent operations across workloads and data stages with differing I/O profiles, and the security and manageability of multitenant environments.

While the data scientists may not care much about the details of the storage infrastructure supporting the AI data pipeline, they do care about the ability of it to gracefully scale to accommodate data growth over time without impacting performance. IDC's *AI DL Training Infrastructure Survey* found that the end users for these workloads (i.e., data scientists, data engineers, and/or software developers) identified provisioning for new projects, scaling existing projects, and the ease and speed of loading data sets for new training runs as key ease-of-use benefits. The ability to deliver consistent performance at scale is important not only to handle data growth but also to accommodate the addition of new applications and users that may want to simultaneously use the same data set. End users also care about the integrity of the data; how easy it is to create copies of data and make it available for use with other applications, to their colleagues, and for rapid recovery purposes; and that the data is protected against failures so that it is not lost. Data scientists also generally prefer tools developed in Python.

Those higher-level objectives translate into a number of specific requirements for the IT administrators actually managing the storage infrastructure for AI training workloads. A primary capability needed in storage that will be used for AI is its ability to support high degrees of concurrency. The same data is generally used across different stages, each of which can have very different I/O profiles, and multiple stages of the AI data pipeline will be operating concurrently much of the time. For example, depending on the scale of data collection, the ingest stage can demand extremely high sequential write performance, while the model training stage typically requires very high random read performance (with a smaller percentage of random writes). For real-time workloads, which are on the rise for enterprises, the production inferencing stage can also require extremely rapid response, driving the need for very low latencies.

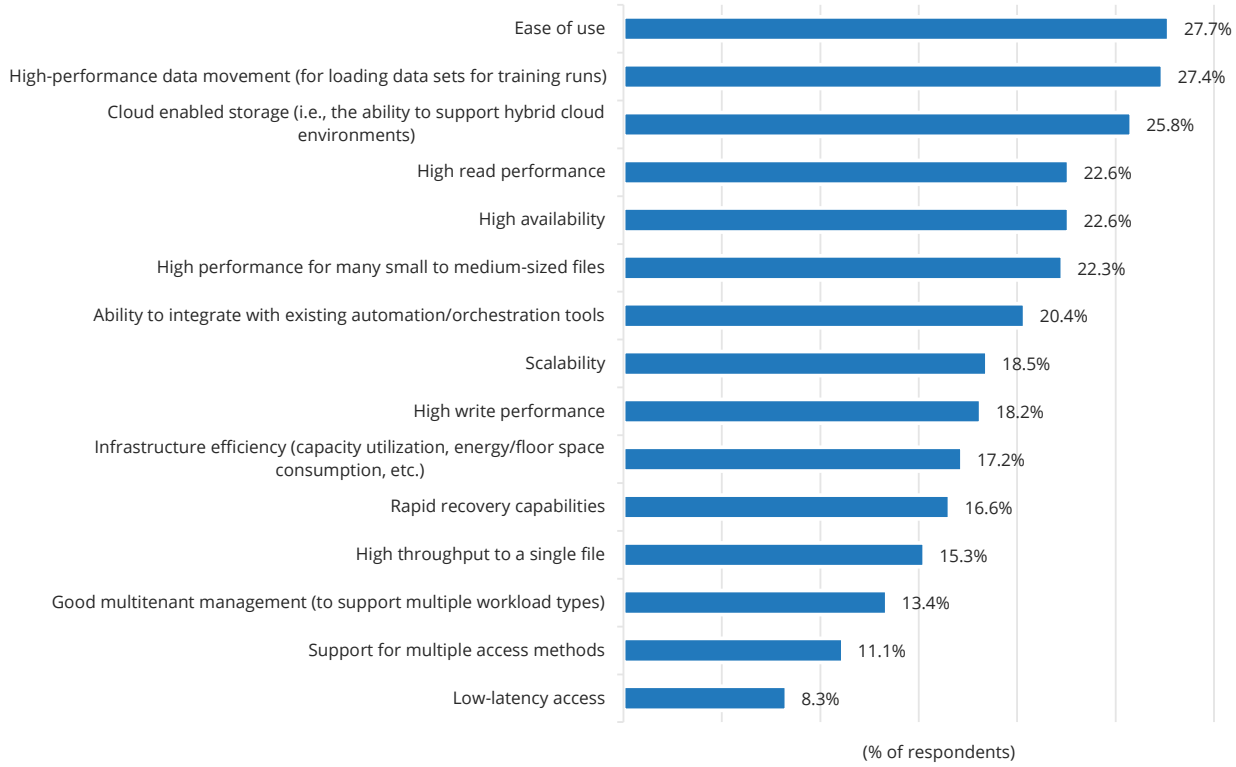
IT administrators were clear about the capabilities topping their list of storage purchase criteria in IDC's *AI DL Training Infrastructure Survey*. Ease of use and high-performance data movement (for rapidly loading data sets for new training runs) were in a virtual tie at the top, followed by cloud-enabled storage, high read performance, and high availability. For those enterprises running multiple AI workloads on the same storage system and sharing data across applications, one of the driving forces behind this workload consolidation was a significant reduction in the time spent moving data *between* different storage systems that were hosting various stages of the AI data pipeline. 52% of respondents were either already consolidating AI workloads onto fewer platforms or expressed an interest in moving to higher levels of consolidation, and 75% are prioritizing data sharing without data movement (a factor often requiring support for higher-speed host networks like NVMe over Fabrics). The callout for high availability is driven by the increasing importance of AI workloads and their criticality to daily business operations that would suffer in the event of outages.

Figure 4 lists storage infrastructure criteria that enterprises are looking at to support AI deep learning training workloads.

FIGURE 4

AI DL Training Storage Purchase Criteria

Q. What features are most important in selecting storage infrastructure to support AI deep learning training workloads?



n = 314

Base = all respondents

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

Enterprises expressed a strong preference for mixed media support in the storage systems used for AI training workloads. Storage administrators need to be able to field all-flash, hybrid, and hard disk drive (HDD)-based nodes to meet different performance, capacity, and cost requirements, and they need the ability to easily accommodate new storage device types as they become available. While HDD storage can provide high capacity at a low price per gigabyte for archive and other less performance-sensitive AI data stages, all-flash configurations were used to support low latency, high degrees of concurrency, higher storage density (i.e., TB/U), and better infrastructure efficiency for performance-sensitive data stages and workloads. NVMe-based flash is particularly important in making efficient use of GPUs since they operate with much higher performance than general-purpose CPUs in data-intensive environments. IDC would also note that for data services like compression, deduplication, and encryption that run inline, flash storage provides a much better ability to maintain low-latency performance over time as a system grows.

When it comes to high-level architectural considerations, working with two-tiered storage infrastructure was strongly preferred. 90% of survey respondents were working with a file system–based front end, which tiered data to a back-end object-based storage platform. For many enterprises, the front end is all-flash while the back end is HDD based. For 42% of enterprises, that object storage platform was on premises, while for 48% it was in the public cloud. Enterprises like the two-tiered approach as it helps keep the front-end file system operating more efficiently and with higher performance. IT administrators want systems that not only enable this capability across potentially multiple cloud targets but also want data migration that is easy to set up and designed to operate efficiently across wide area networks. Mature hybrid cloud integration helps make this easy and is one of the factors driving the broad usage of hybrid cloud infrastructure.

Enterprise Deployment Model Preferences

Simplified deployment models are very important to enterprises as they offer easier ordering, quicker installation, faster time to value and, in many cases, streamlined support. Converged infrastructure offerings, which bundled compute, storage, and networking into a factory-configured rack that could be purchased under a single SKU and offered a single point of support contact for all infrastructure components, were introduced in the early 2010s and have grown into a $21.3 billion market (in 2021). This same idea has taken hold in the AI infrastructure market, and there are now several enterprise storage providers that have created converged infrastructure stacks that are specifically targeted at AI workloads. Mostly, these systems leverage NVIDIA accelerated compute and networking, and vendor-specific storage, and generally, in addition to the benefits previously mentioned, also provide a unified management interface that further simplifies the management of these environments.

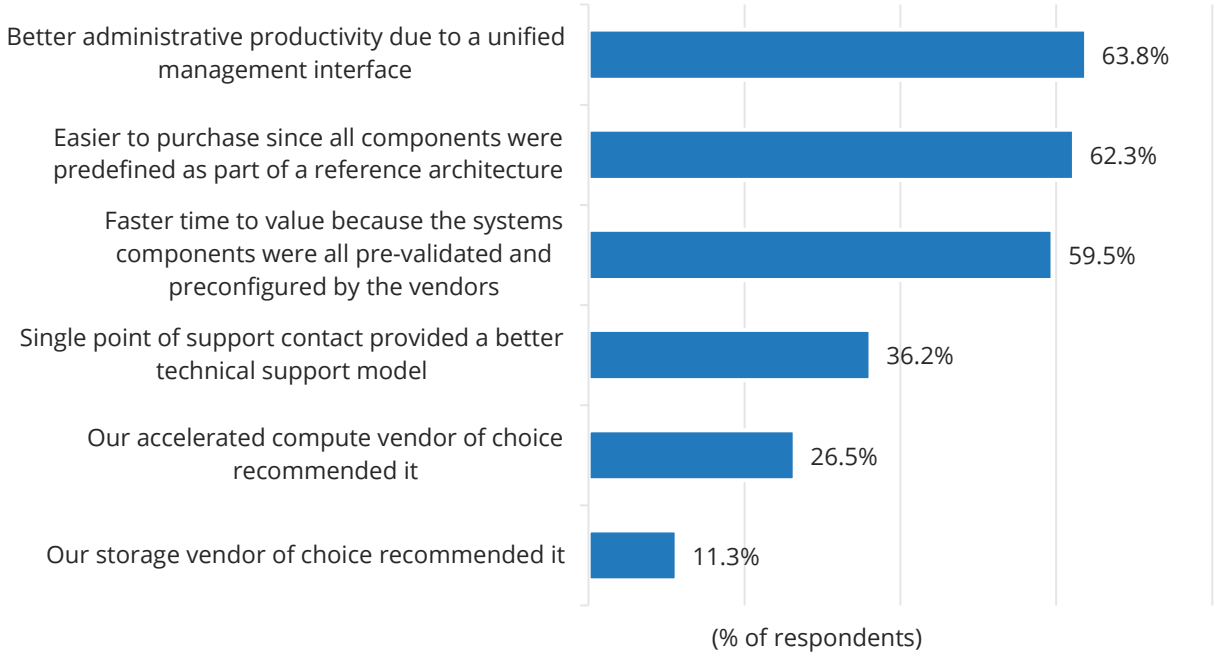
IDC's *AI DL Training Infrastructure Survey* found that more than 83% of the enterprises that had purchased a storage system for AI workloads had also at some point purchased a converged infrastructure offering. The top drivers of those purchases included better administrative productivity (due to the unified management interface), ease of purchasing based on the implied reference architecture, a faster time to value, and a single point of support contact. When converged infrastructure stacks already include most of the components an enterprise would have bought separately anyway, they are a much better way to go. And even with these offerings, enterprises do have some flexibility in adding other components they may need going forward (although those may or may not be pretested by vendors and fall under the single support contract).

Figure 5 lists the major reasons behind enterprises' converged infrastructure stack purchase decisions for AI DL training workloads.

FIGURE 5

Drivers of Converged Infrastructure Stack Purchases for AI DL Training Workloads

Q. What were the reasons behind your converged infrastructure stack purchase decision?



n = 257

Base = respondents indicated organization purchased converged infrastructure stack for AI training workloads

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

Reference architectures are available from many vendors as well, although their benefits do not go quite as far as converged infrastructure stacks. A reference architecture specifies a pretested configuration using multivendor components so that they have been validated to work together but leaves it up to the customer to buy the components from their various vendors. Ordering a complete system requires more manual effort on the part of customers, support contacts are split across the various vendors, and there is no unified management interface. When converged infrastructure stacks that include the products an enterprise wants are not available, it may be better to work from a reference architecture since enterprises will not need to validate all the product combinations themselves.

NETAPP SOLUTIONS FOR AI-DRIVEN WORKLOADS

NetApp is a $6.3 billion hybrid cloud data services and data management vendor headquartered in San Jose, California. Today, the vendor is recognized as a leader in enterprise storage, and its broad portfolio includes block-, file-, and object-based storage platforms as well as converged infrastructure, technical support, and consulting services – all based around a software-defined product strategy that offers outright purchase and subscription-based deployment options. As the first major enterprise

storage vendor to recognize how important hybrid cloud infrastructure would become to IT, it began building out its hybrid cloud portfolio in 2015 under new (at the time) CEO George Kurian. Today, NetApp has an extremely mature hybrid cloud offering, allowing customers to run their enterprise-grade solutions either on premises or in major hyperscaler public cloud environments (AWS, Microsoft Azure, Google Cloud, etc.) and manage them all under a unified management interface that supports hybrid multicloud operations.

Working jointly with Cisco back in 2010, NetApp created the converged infrastructure market with its FlexPod offering. In 2018, NetApp partnered with NVIDIA to create a converged infrastructure stack for enterprise AI called NetApp ONTAP AI. The system combined NVIDIA DGX accelerated compute systems and networking, and NetApp NVMe-based all-flash storage (the NetApp AFF A800 at the time), along with a unified management interface, single point of support contact, and integration with the NetApp Data Fabric. Through the NetApp Data Fabric, the vendor's unifying architecture to standardize data management practices across cloud, on premises, and edge devices, NetApp ONTAP AI enables enterprises to create a seamless data pipeline that spans from the edge to the core to the cloud for AI workloads. The benefits of this targeted solution for enterprise AI workloads included fast and easy ordering and deployment, an ability to nondisruptively scale to hundreds of petabytes of cluster capacity, and the ability to operate with confidence using NetApp's enterprise-grade, highly available storage platforms.

The vendor has kept this converged infrastructure offering up to date as new accelerated compute and storage platforms have become available. Today's NetApp ONTAP AI combines the NVIDIA DGX A100, the world's first 5PFLOPS system, with the new NVMe-based NetApp AFF systems and the latest NVIDIA high-performance networking. The NVIDIA DGX A100 has the power to unify AI workloads for training and inferencing as well as data analytics and other high-performance workloads on a single compute platform, while the high-end NetApp A900 delivers the low latency, high availability, and scalable storage capacity and the proven enterprise storage capabilities to meet the needs of both data scientists and IT administrators. The converged infrastructure offering is available in 2, 4, and 8-node (referring to the number of interconnected DGX systems) configurations combined with optimized AI software.

Bundled NetApp ONTAP AI solution components include the NVIDIA Base Command software stack, the NetApp AI Control Plane, and the NetApp DataOps Toolkit. The NVIDIA Base Command software stack is a full-stack suite of pre-optimized AI software including a DGX-optimized OS, drivers, infrastructure acceleration libraries that speed I/O, enterprise-grade cluster management, job scheduling and orchestration, and full access to the NVIDIA AI Enterprise software suite for additional developer assets like optimized frameworks, pretrained models, model scripts, AI and data science tools, and industry SDKs. The NetApp AI Control Plane integrates Kubernetes and KubeFlow with the NetApp Data Fabric to simplify data management in multicloud environments, while the NetApp DataOps Toolkit is a Python library that makes it easy for data scientists to perform common data management tasks in AI environments like provisioning new storage, cloning data, and creating snapshots for traceability and baselining. ONTAP AI also supports NVIDIA GPU Direct Storage for its ONTAP NFS and E-Series BeeGFS parallel file system.

At the heart of ONTAP AI are NVIDIA DGX compute systems, a fully integrated hardware and software turnkey AI platform that's purpose built for analytics, AI training, and AI inferencing, delivering 5PFLOPS with a 6U form factor. The NVIDIA DGX A100 integrates eight NVIDIA A100 Tensor Core GPUs, interconnected with the NVIDIA NVSwitch architecture, offering an ultra-high-bandwidth, low-

latency fabric over which AI workloads can be parallelized across multiple GPUs with NVIDIA InfiniBand enabling workloads to be distributed over many DGX systems in an infrastructure cluster.

ONTAP AI Built on NVIDIA DGX BasePOD

Since 2018, NetApp and NVIDIA have served hundreds of customers with a range of solutions, from building AI centers of excellence to solving massive-scale AI training challenges. NVIDIA DGX BasePOD is the reference architecture derived from years of experience and expertise gained from AI infrastructure deployments around the world. ONTAP AI built on NVIDIA DGX BasePOD eliminates the guesswork for faster adoption by using this field-proven blueprint for scaling compute, storage, and networking in an AI infrastructure. It's a preconfigured, integrated solution that's easy to procure and offers turnkey deployment. NetApp is the only NVIDIA DGX BasePOD partner that runs the same data fabric on premises and in the public cloud.

NVIDIA DGX A100 features NVIDIA ConnectX-7 InfiniBand/Ethernet network adapters with 500GBps (gigabytes per second) of peak bidirectional bandwidth. These network adapters enable the DGX A100 to serve as the building block for large AI clusters such as NVIDIA DGX SuperPOD, enabling datacenter-scale AI performance and supporting datacenter-scale workloads.

The storage in NetApp ONTAP AI is based around the vendor's flagship scale-out storage operating system (ONTAP). Of interest to data scientists, ONTAP environments can nondisruptively scale from 2 nodes up to 24 nodes in a single cluster and support over 700PB across multiple namespaces (a single namespace can host over 20PB of raw capacity). Provisioning new storage is intuitive, fast, and very easy through the NetApp Cloud Manager. The use of end-to-end NVMe technology allows NetApp ONTAP AI clusters to support very low latencies and a high degree of concurrency to support multiple applications simultaneously working with data across multiple data pipeline stages and support very high-speed data loading.

NetApp worked with NVIDIA to produce several reference architectures, based on ONTAP AI, that are targeted for use cases in different industries. Available NetApp ONTAP AI reference architectures include:

- ONTAP AI Reference Architecture for Healthcare: Diagnostic Imaging
- ONTAP AI Reference Architecture for Autonomous Driving Workloads: Solution Design
- ONTAP AI Reference Architecture for Financial Services Workloads: Solution Design
- ONTAP AI Reference Architecture for Sentiment Analysis for Text and Audio

All ONTAP storage systems come with high availability, storage efficiency, data management, scalable NAS data protection, security, compliance, and cloud integration to ensure data integrity. Built as they are for mission-critical workloads, the AFF systems transparently recover from a variety of different failures, support "hot" replacement of failed components, and include replication-based disaster recovery technology that ensures that data scientists' data will always be accessible. Data copy creation is also fast and easy, using space-efficient delta differential technologies to speed copy creation and maximize capacity utilization. File-level recoveries use NetApp's snapshot technology that will in many cases access co-located copies within the storage array itself to perform data recovery in literally seconds.

Storage administrators will also appreciate the comprehensive and proven enterprise storage management feature set on all ONTAP arrays. ONTAP is an enterprise-class, clustered data management solution that delivers high performance, high capacity, nondisruptive operations (to

support the vendor's 100% data availability guarantee), and comprehensive enterprise-class data services enabling secure multi-tenancy, tiered storage configurations (both within systems and across hybrid multicloud environments), deep enterprise application and cloud integration, and a wide range of automated operations for common datacenter workflows that leverage ONTAP's REST API.

ONTAP runs on both NetApp's NVMe-based all-flash systems (AFF) and its hybrid and/or HDD-based systems (FAS) and supports unified (block and file) as well as block-only (all-SAN array) storage. Supported access methods for ONTAP include NFS, SMB, and S3 as well as Fibre Channel (FC) and iSCSI (for block access). It includes a wide range of storage efficiency technologies, space-efficient snapshot technology (including immutable snapshots), and many replication options (including both synchronous and asynchronous as well as Metro Clusters for instant recovery with zero data loss in metro environments) and support for popular APIs from vendors such as VMware, Microsoft, and Oracle. Storage administrators interested in a deeper take from IDC on what ONTAP brings to the table can refer to *Meeting the High Availability Requirements in Digitally Transformed Enterprises* (IDC #US48442021, March 2022).

ONTAP also supports tiering to external S3 targets whether those are on premises (e.g., using the vendor's object-based storage called StorageGRID) or off premises (any public cloud provider that supports S3 storage). The use of external storage targets can provide a very cost-effective, massively scalable platform for long-term archiving. It also can provide a potential secondary HDD-based tier (with the first HDD-based tier being in ONTAP-based FAS systems).

ONTAP also supports S3 protocol access to allow NFS mount points and associated files to be accessed as buckets via the S3 protocol — a feature named multiprotocol access.

When Scaling Beyond NFS Is Required

NetApp ONTAP AI enables significant consolidation of different stages of the AI data pipeline onto a single ONTAP-based system and with its low latency enables very high utilization of accelerated compute resources. There are cases, however, when workload requirements demand more throughput than is possible with NFS. For these environments, customers often need an intelligent (parallel) client that can provide tens to hundreds of gigabytes per second in throughput to very large files.

BeeGFS is a parallel file system with a distributed metadata architecture that was developed and optimized for HPC environments demanding extremely high throughput to very large files. It is POSIX compliant, offers an extremely scalable intelligent (parallel) client, and is widely used with technical and HPC workloads. NetApp has created a reference architecture around BeeGFS called "NetApp E-Series Storage with BeeGFS" to meet these types of requirements, which can also arise with certain AI-driven workloads. This reference architecture is based on the NetApp E-Series systems, which support all-flash, hybrid, and all-HDD-based systems to meet a variety of requirements. Included within the E-Series portfolio are the EF-Series systems, which are all-flash, based on NVMe, and deliver the lowest latencies within the NetApp storage portfolio for the most demanding workloads. The E-Series systems are enterprise grade, include a variety of storage management capabilities, support Ansible automation, and feature "six nines" (99.9999%) availability in a highly scalable system built around performance-optimized storage building blocks. Although BeeGFS is free open source software, NetApp customers purchasing this solution can obtain a support contract that comprehensively covers both the hardware and software in this reference architecture solution.

When customers are looking for a parallel file system, prefer to use InfiniBand for storage interconnect, and are working with single long-running jobs using larger file sizes, NetApp E-Series Storage with

BeeGFS is a good fit. Big data applications and workloads in the oil and gas, genomics, and complex simulation arenas drive the need for a parallel file system like BeeGFS, along with technical computing, HPC, and large-scale natural language processing environments.

NetApp Keystone: Delivering a Consistent Cloud Experience Everywhere

One of the things enterprises liked most about the public cloud model was the ability to align the payment for IT resources with the usage of the resources. While some customers still prefer up-front capex or leasing, having subscription-based licensing as an option is becoming very important for digitally transforming enterprises evolving their hybrid cloud infrastructures. Keystone is NetApp's pay-as-you-grow, storage-as-a-service offering that includes a unified management console and consolidated monthly billing for both on-premises and cloud-based services. It offers a number of different storage services options including on premises, self-managed, or managed (by NetApp) infrastructure and managed services at colocation sites (e.g., Equinix). NetApp Keystone applies to all of the NetApp storage systems products discussed in this white paper. Those include all NetApp A Series systems (AFF 250, 400, 800, and 900), all FAS systems (FAS27XX, FAS 87XX, FAS9000, and FAS500f), NetApp E-Series Storage, and StorageGRID (all SG and SGF models). Converged infrastructure stacks like NetApp ONTAP AI are also covered under Keystone, although those systems all use NetApp AFF storage.

NetApp Cloud Storage

NetApp has optimized the ONTAP storage system to run in the public clouds. ONTAP is offered as a managed first-party public cloud service from the major hyperscalers (e.g., Amazon FSx for NetApp ONTAP, Azure NetApp Files). NetApp has really distinguished itself with its ability to partner with the hyperscalers to deliver first-party storage services offerings that offer a better integrated enterprise storage experience from service providers than just licensing storage software to deploy on a service like AWS EBS or Azure Virtual Disks. With better performing storage for AI in the cloud, NetApp storage allows AI training jobs to run faster than with native cloud storage.

NetApp has delivered solutions in the cloud for AI use cases, including:

- Lane Detection with Azure NetApp Files
- Azure Machine Learning with Azure NetApp Files
- Distributed training in Azure — Click-Through Rate Prediction with Azure NetApp Files

CHALLENGES/OPPORTUNITIES

Enterprises vary in terms of who makes the buying decisions for AI infrastructure. IDC's *AI DL Training Infrastructure Survey* indicated that the CIO/CTO level made that decision in almost 40% of cases, while the IT infrastructure team made it in 34% of cases and lines of business made it in over 10% of cases. Data scientists and AI application developers made that decision in 9.5% of cases. To select the best system for a given enterprise's AI training workloads, decision makers must have a good idea of the requirements from the data scientist/developer side as well as the IT operations side. When selecting the right storage infrastructure for these workloads, it is important that all those affected are consulted and have an opportunity to agree on objectives and priorities.

The opportunity for enterprises when adding a new storage system specifically for AI workloads is to determine how much other workload consolidation that platform could (and should) support. The more AI data pipeline stages can be consolidated onto a single storage platform, the better, and when that

platform has the performance and scalability to support additional workloads beyond AI, enterprises can reap a great return on investment. But AI projects are usually under a tight schedule, and these types of considerations can lengthen the time to deploy. Vendors like NetApp whose systems are high performance and highly scalable can allow enterprises to use workload consolidation to drive very high infrastructure efficiencies. These types of vendors have an opportunity to help enterprises understand their objectives and the requirements of different constituencies up front to help make a quick, AI workload–enabling storage decision.

Within enterprises themselves, it is to their benefit if the requirements of all constituencies are understood up front as well, so they can be very clear with potential storage vendors what their different organizations require in an AI storage platform. Enterprises should seek to maximize workload consolidation during infrastructure modernization efforts to drive better infrastructure efficiencies but need to ensure that they do not compromise performance, availability, or security in doing so.

CONCLUSION

While it is still early in the move toward AI in the enterprise, it is clear that this will be a central and strategic workload in digitally transformed enterprises. Storage infrastructure spend in enterprises for AI workloads alone will be a $5.4 billion market by 2024. A large percentage of this spend will be for the replacement of existing storage systems that, after getting past AI pilots, enterprises discover cannot provide the performance and scalability needed for AI workloads. DL workloads in particular require large data sets and delivering performance consistently at scale for those applications outpaces the capabilities of many general-purpose storage systems.

AI workloads generally have a number of data pipeline stages running from ingest and transformation through training, inferencing, and archiving, and enterprises are rightly focusing on storage systems that can simultaneously handle all those stages at once without undue performance and availability impacts. When a single storage system can meet all the requirements, it offers far better economics than older approaches that require storage silos for one or more stages. As enterprises go through digital transformation in general, they are looking to improve IT infrastructure efficiencies, and being able to consolidate all stages of AI workloads on a single system goes a long way toward achieving that goal. When those storage systems have the performance and capacity to be able to consolidate additional workload types, the economic advantages only get better.

NetApp ONTAP AI is a converged infrastructure stack that combines NVIDIA DGX accelerated compute, NVIDIA high-speed switching, ONTAP-based storage, and a large selection of tools to help manage AI workloads effectively. The stack combines all of those components into an integrated system, purchased under a single SKU, that is easy to buy and deploy, is fully supported by NVIDIA, and includes a unified management interface that boosts administrative productivity for these types of configurations. These systems are based on proven technologies, and NetApp has deployed hundreds of these systems for AI workloads in enterprises over just the past several years. Overall, NetApp has tens of thousands of storage systems in enterprises across all workloads and use cases, but the vendor has been very successful in delivering rapid time to value for customers using ONTAP for AI workloads. Almost 60% of enterprises are also running non-AI workloads on the storage systems they are using for those workloads, and NetApp ONTAP offers a comprehensive set of capabilities to enable these types of multitenant environments without compromising performance, availability, or

security. Enterprises in the process of adding and/or scaling AI-driven workloads would do well to consider NetApp as the storage platform for these environments.

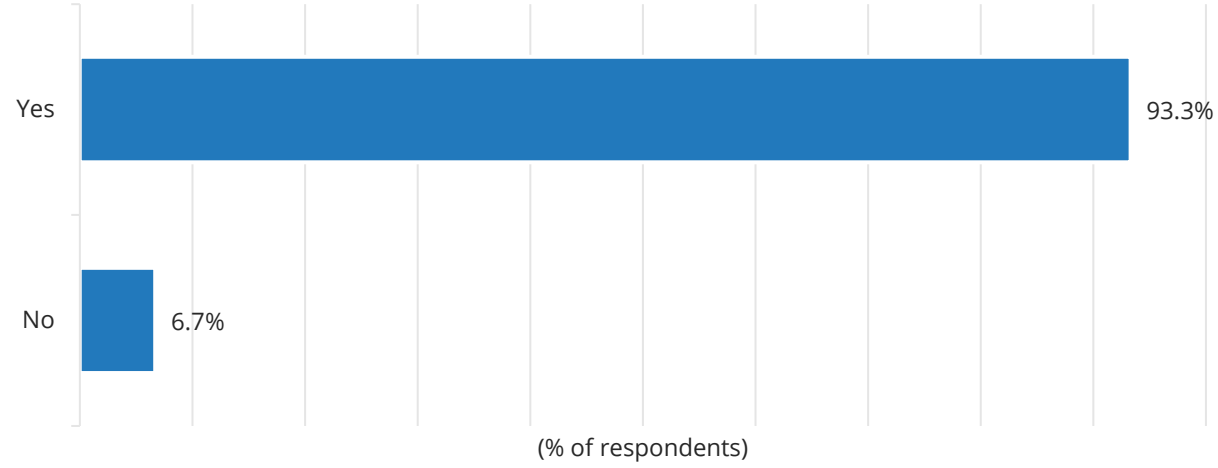
APPENDIX

Figures 6-9 provide additional results from IDC's *AI DL Training Infrastructure Survey.*

FIGURE 6

AI DL Training Workloads Will Remain Popular Over the Next Year

Q. Are you thinking of running AI deep learning training workloads in your IT environment within the next 12 months?



n = 314

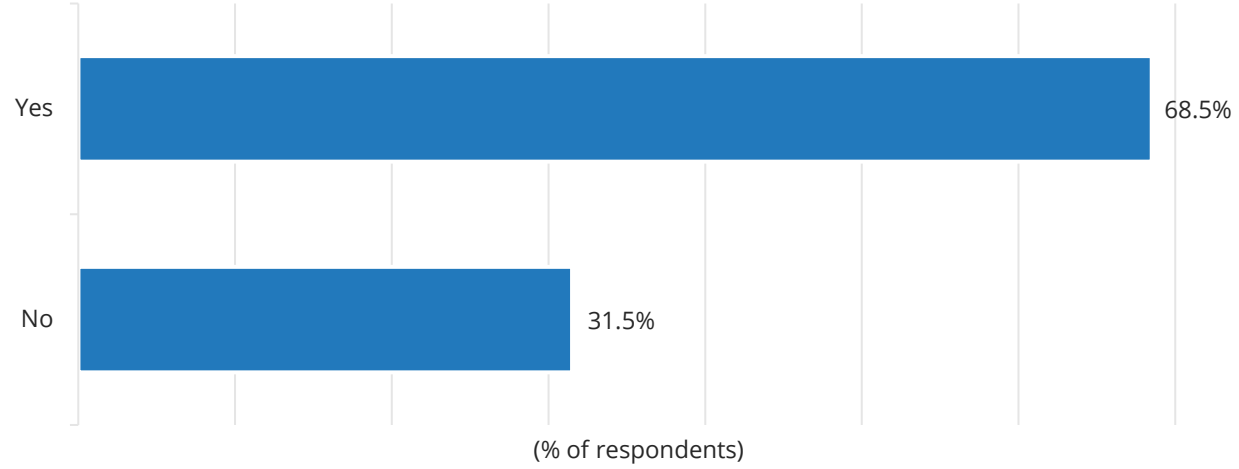
Base = all respondents

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

FIGURE 7

69% Move AI DL Data Between On Premises and Public Cloud

Q. Are you moving data between on premises and public cloud locations?



n = 251

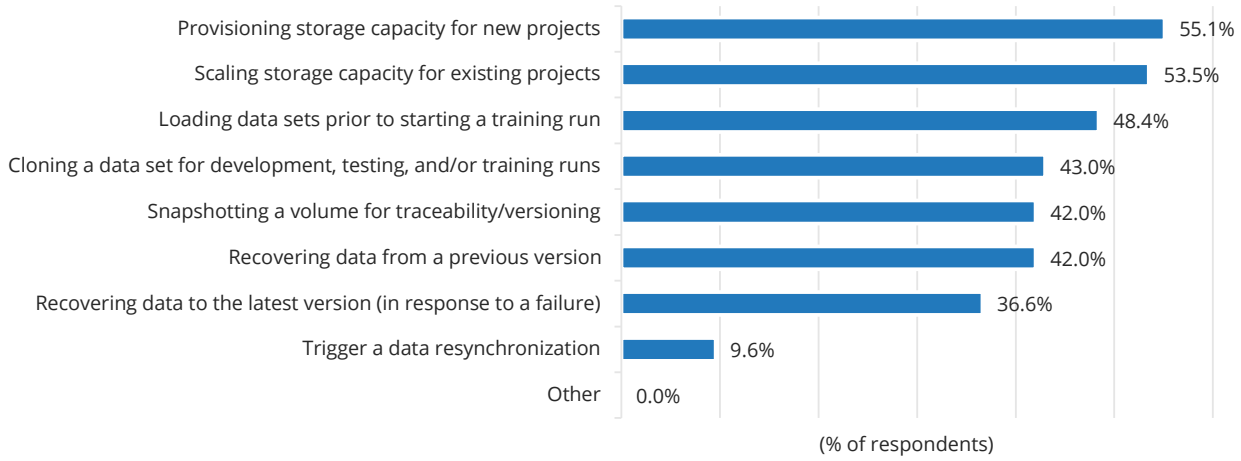
Base = respondents indicated organization currently running AI deep learning training projects in on-premises only/a nonintegrated mix of on premises and the cloud/an integrated hybrid cloud.

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

FIGURE 8

Top Storage Ease-of-Use Benefits for End Users — Provisioning, Scaling

Q. Which of the following are the core capabilities your AI deep learning training storage infrastructure needs to make it easily accessible for data scientists, data engineers, and/or software developers?



n = 314

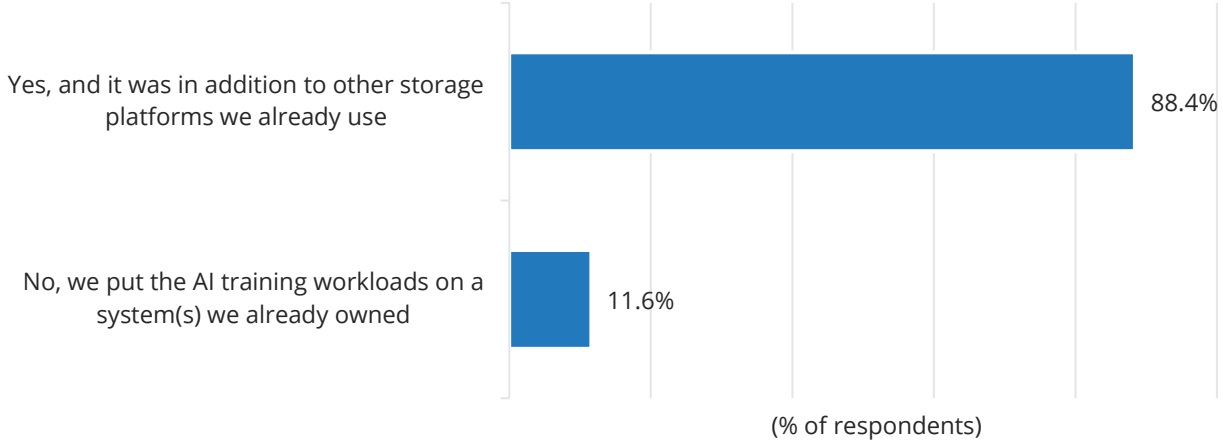
Base = all respondents

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

FIGURE 9

Users Buy a Storage System Specifically for AI DL Training Workloads

Q. Did you buy a storage system specifically for your AI deep learning training workloads?



n = 294

Base = respondents indicated organization currently running AI deep learning training workloads in IT environment

Source: IDC's *AI DL Training Infrastructure Survey,* June 2022

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