insideBIGDATA

insideBIGDATA Guide to

Optimized Storage for AI and Deep Learning Workloads

By Daniel D. Gutierrez, Managing Editor & Resident Data Scientist, insideBIGDATA

BROUGHT TO YOU BY DDN® AI · BIG DATA · HPC

Contents

Introduction

Artificial Intelligence (AI) and Deep Learning (DL) represent some of the most demanding workloads in modern computing history as they present unique challenges to compute, storage and network resources.

Traditional file storage technologies and protocols like NFS restrict AI workloads of data, thus reducing the performance of applications and impeding business innovation. A state-of-the-art AI-enabled data center should work to concurrently and efficiently service the entire spectrum of activities involved in DL workflows, including data ingest, data transformation, training, inference, and model evaluation.

Optimized storage has a unique opportunity to become much more than a siloed repository for the deluge of data constantly generated in today's hyper-connected world, but rather a platform that shares and delivers data to create competitive business value. Optimized storage is designed for the needs of a broad range of problem domains including FinTech, life sciences, design, HPC, government, smart cities, media, energy, and many more.

Optimized storage solutions are faced with I/O intensive workflows found in AI. Traditional structured data infrastructure is acceptable in many cases, whereas an optimized storage platform is targeted at other classes of problems that tend to be unstructured and throughput intensive.

Today, we're seeing an emphasis for rebalancing the cloud. AI is bucking the trend toward migration to the cloud. By their nature, some enterprises generate a lot of unstructured data at the edge, e.g. in a vehicle, in a store, from a microscope in a lab. These use cases are very non-cloud-friendly since there is a need to ingest a large amount of data. With technologies like 5G and IoT, it's important to do real-time analytics on that data. This means huge data volumes come into play, and real-time analytics is changing what was previously seen as a homogeneous move of everything to the cloud. That emphasis is being stopped in its tracks now, and there is much more of an equilibrium state of edge and on-prem gains with a change in IT thinking these days. It's not just an "all cloud" story, but a more complex narrative.

The intended audience for this important new technology guide includes enterprise thought leaders (CIOs, director level IT, etc.), along with data scientists and data engineers who are a seeking guidance in terms of infrastructure for AI and DL in terms of specialized hardware. The emphasis of the guide is "real world" applications, workloads, and present day challenges.

How Optimized Storage Solves AI Challenges

The IT infrastructure supporting an AI-enabled data center must adapt and scale rapidly, efficiently and consistently as data volumes grow and application workloads become more intense, complex and diverse. The nature of DL deployments means IT resources must seamlessly and continuously handle transitions between different phases of experimental training and production inference in order to provide faster and more accurate answers. In short, the IT infrastructure is instrumental to realizing the full potential of AI and DL in business and research.

In order for GPUs to fulfill their promise of acceleration, data must be processed and delivered to the underlying AI applications with great speed, scalability and consistently low latencies.

Yet as demand continues to rise, current enterprise and research data center IT infrastructures are sadly inadequate in handling the challenging needs of AI and DL. Designed to handle modest workloads, minimal scalability, limited performance needs and small data volumes, these platforms are highly bottlenecked and lack the essential capabilities required for AI-enabled deployments.

Data storage is a central area of focus. There are a number of easily identified differences between traditional storage and optimized AI data platforms. For example, challenges with traditional storage include: low speed, poor latency, no graphics processing unit (GPU) integration, no container optimization, limited scaling, no multiple writers, and inefficient TCP/IP communication. In contrast, AI data platform benefits include: fully saturates GPU/CPU, maximizes efficiency at scale, continuous data availability, highest deep learning acceleration, seamless scalability, effortless deployment and management.

Optimized storage platforms for AI and DL workloads offer support for a broad range of uses cases in the following ways:

- Accelerate applications by achieving full GPU saturation
- Streamline concurrent and continuous DL workflows
- Flexible configuration with best technology and economics
- Seamless scaling to match evolving workflow needs

Parallel Data from Storage to GPU

Revolutionary breakthrough technologies in processors and storage serve as important catalysts for effective AI data center enablement. For example, GPUs deliver compute acceleration over slower CPUs, while Flash Enabled Parallel I/O Storage provides a significant performance boost to legacy hard disk-based storage. Specifically, GPUs are significantly more scalable and faster than CPUs while their large number of cores permit massively parallel execution of concurrent threads. This parallelism results in accelerated training and inference capabilities for AI/DL applications.

In order for GPUs to fulfill their promise of acceleration however, data must be processed and delivered to the underlying AI applications with great speed, scalability and consistently low latencies. This requires a parallel I/O storage platform for performance scalability, real time data delivery, and flash media for speed. Without the right data storage platform, a GPU-based computing platform is just as bottlenecked and lacking as a traditional non-AI-enabled data center. The proper selection of the data storage platform and its efficient integration in the data center infrastructure are key factors to eliminating AI blockages and truly accelerating time-to-insight.

The right data storage system must deliver high throughput, high IOPS and high concurrency in order to prevent idling of precious GPU cycles. It must be flexible and scalable in implementation and enable efficient handling of a wide breadth of data sizes and types, including highly concurrent random streaming, a typical DL data set attribute.

Properly selected and implemented, such a data storage system will deliver the full potential of GPU computing platforms, accelerate time-to-insight at any scale, effortlessly handle every stage of the AI and DL process, and do so reliably, efficiently and cost effectively.

GPUs provide a powerful platform for AI. Their high number of cores deliver a massive parallel computing facility that can process very large amounts of data simultaneously. To achieve the full potential of AI and DL applications and maximize the benefits of GPUs, data saturation of all cores must be achieved and sustained. Fulfilling this requirement for multiple GPUs simultaneously poses a significant technical challenge.

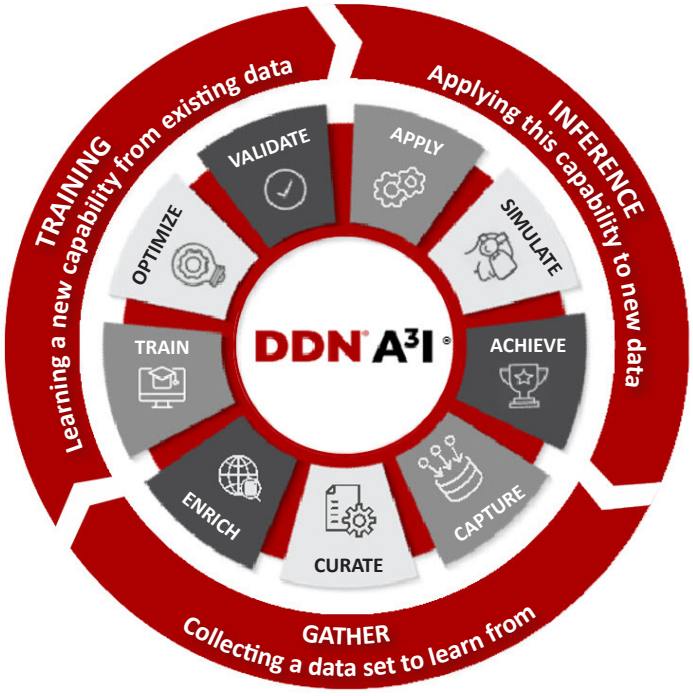
It's necessary to provide a highly scalable shared storage platform which effortlessly integrates with multi-GPU computing environments, while maintaining concurrent data saturation. The platform can start small while still delivering the performance needed for GPU saturation, and scale seamlessly in performance, capacity and capability. As data sets grow and additional GPUs are deployed, it's critical to continuously deliver an optimized, extremely cost-effective solution. The parallel architecture and protocol should deliver data with high-throughput, low-latency, and massive concurrency. It should also provide increased performance for DL frameworks and offer significantly faster processing than NFS.

Optimized AI Infrastructure in Support for Real-world Applications

Many organizations have become more entrenched in running real world workloads and real world applications based on AI and DL. Here are some examples of proven ways optimized storage solves AI challenges:

- **AI powered retail –** deploy and manage data solutions across thousands of remote, autonomous checkouts
- **Financial services –** demonstrate due diligence and adherence to regulations around customer data
- **Language processing –** guarantee a real-time experience during peak loads
- **Autonomous vehicles –** manage hundreds of petabytes of globally distributed data
- **Life sciences and healthcare –** securely extract value from patient data in real-time
- **HPC –** integrate HPC with AI, and accelerate tough HPC and AI workloads

Optimized storage solutions work to effortlessly address the complete AI lifecycle as shown in the figure below. The AI workflow is fully optimized concurrently, continuously, and in-place.



A³I – Accelerated, Any-scale AI Solutions

DDN's A³I® (Accelerated, Any-Scale AI) optimized storage solutions break into new territory for AI and DL - enabling platforms that can maximize business value in terms of an optimized AI environment: applications, compute, containers, and networks.

Engineered from the ground up for the AI-enabled data center, A³I solutions provide acceleration for AI applications and streamlined DL workflows using the DDN shared parallel architecture. Flash can scale-up or scale-out independently from capacity layers, all within a single integrated solution and namespace. A³I solutions provide unmatched flexibility for your organization's AI needs.

DDN A³I solutions are easy to deploy and manage, highly scalable in both performance and capacity, and represent a highly efficient and resilient platform for your present and future AI requirements. Below is a summary of the benefits of A³I solutions along with product summaries for the DDN AI200, AI400, AI7990.

Benefits & Product Summaries

- Delivers a fully integrated and optimized data flow from application to flash and disk that ensures full saturation of GPUs for maximum productivity at any scale.
- Offers the best performance, efficiency and economics for computation, training, inference and data transformation.
- Provides high capacity AI storage - up to 240TB of reliable and efficient dual ported NVMe flash in a highest-density 2RU configuration. The solutions can scale seamlessly to multiple petabytes of flash as workflow needs evolve to provide additional capacity, performance and capability.
- Offers high resiliency, reliability, security at scale. The solutions are engineered to provide highest data availability and maximum system uptime.
- The solutions integrate a true distributed parallel file system that provides a very simple highly scalable, single namespace structure.

DDN AI200® and AI400®

DDN's AI200 and AI400 are efficient, reliable and easy to use data storage systems for AI and DL applications. With end-to-end parallelism, the AI200 and AI400 eliminate the bottlenecks associated with NFS-based platforms and deliver the performance of NVMe Flash directly to AI applications.

Extensive testing with widely-used AI and DL applications demonstrates that a single AI200 provides tremendous acceleration for data preparation, neural network training and inference tasks using GPU compute servers.

Both the AI200 and AI400 can scale horizontally as a single, simple namespace. They integrate tightly with hard disk tiers to help manage economics when data volumes expand. The AI200 and AI400 are specifically optimized to keep GPU computing resources fully saturated, ensuring maximum efficiency while easily managing challenging data operations, from non-continuous ingest processes to large scale data transformations.

DDN's A³I reference architectures are designed in collaboration with NVIDIA® and HPE to ensure highest performance, optimal efficiency, and flexible growth for NVIDIA DGX-1™ servers, NVIDIA DGX-2™ servers and HPE Apollo 6500 servers. Extensive testing with widely-used AI and DL applications demonstrates that a single AI200 provides tremendous acceleration for data preparation, neural network training and inference tasks using GPU compute servers.

Delivering flash performance direct to your AI application, the AI7990 brings new levels of simplicity and flexibility to help deal with unforeseen AI deployment challenges.

DDN AI7990

DDN's AI7990 is a hybrid flash and hard disk data storage platform designed for AI and DL workflows. With end-to-end parallel access to flash and deeply expandable hard disk drives (HDDs) storage, the AI7990 outperforms traditional NAS platforms and delivers the economics of HDD for expanding data repositories. Delivering flash performance direct to your AI application, the AI7990 brings new levels of simplicity and flexibility to help deal with unforeseen AI deployment challenges. AI7990 is faster, denser, more scalable and more flexible in deployment than other storage platforms with up to 23GB/s of filesystem throughput and over 750K IOPS in 4RU as a single namespace to meet demand. DDN flash and spinning disk storage is integrated tightly within a single unit for up to 1PB in just 4RU. As with the AI200/AI400 systems, the AI7990 keeps GPU servers saturated.

A³I with HPE Apollo 6500™

The DDN A³I scalable architecture fully-integrates HPE Apollo 6500 servers with DDN AI all flash parallel file storage appliances to deliver fully-optimized, end-to-end AI and DL workflow acceleration. A³I solutions greatly simplify the use of the Apollo 6500 server while also delivering performance and efficiency for full GPU saturation, and high levels of scalability.

The HPE Apollo 6500 Gen10 server is an ideal AI and DL compute platform providing unprecedented performance supporting industry leading GPUs, fast GPU interconnect, high bandwidth fabric, and a configurable GPU topology to match workloads.

The following reference architectures are specifically designed for AI and DL. They illustrate tight integration of DDN and HPE platforms.

Every solution has been designed and optimized for Apollo 6500 servers, and thoroughly tested by DDN in close collaboration with HPE.

- Configuration for single Apollo 6500 server with AI200
- Configuration for single Apollo 6500 server with AI400
- Configuration for four Apollo 6500 servers with AI200
- Configuration for four Apollo 6500 servers with AI400

Performance testing on the DDN A³I architecture has been conducted with synthetic throughput and IOPs testing applications, as well as widely-used DL frameworks. The results demonstrate that using the A³I intelligent client, containerized applications can engage the full capabilities of the data infrastructure, and that the Apollo 6500 server achieves full GPU saturation consistently for DL workloads.

Delivering flash performance direct to your AI application, the AI7990 brings new levels of simplicity and flexibility to help deal with unforeseen AI deployment challenges.

EXA5 (EXAScaler® Storage)

EXA5 provides deeper, more powerful integrations into AI and HPC ecosystems through simpler implementation and scaling models, easier visibility into workflows and powerful global data management features.

EXA5 allows for scaling either with all NVMe flash or with a hybrid approach, offering flexible scaling according to needs with the performance of flash or the economics of HDDs. Compatible with DDN's A³I storage solutions, which are factory pre-configured and optimized for AI, deployment time is reduced for an AI-ready data center. Together they provide full benchmarking, documentation and qualification with AI and GPU environments.

A primary feature of EXA5 is STRATAGEM, DDN's unique integrated policy engine. STRATAGEM enables platforms for transparent flash tiering

for high performance and efficiency. Featuring built-in fast and efficient data management, users can manage most active data to scale-out flash tiers from scale-out HDD tiers. It also automatically controls free space on flash, quickly responding to changing demands and efficiently scanning storage devices directly to find target files. Other STRATAGEM features include:

- Built-in fast and efficient data management
- Manage most active data into scale-out Flash
- Automatic control of free space on Flash
- Quickly respond to changing demand
- Efficient namespace scanning

EXA5 also includes small file performance enhancements (useful for the mixed workloads of unstructured data), along with integrated end-to-end data integrity verification which verifies data from application to media.

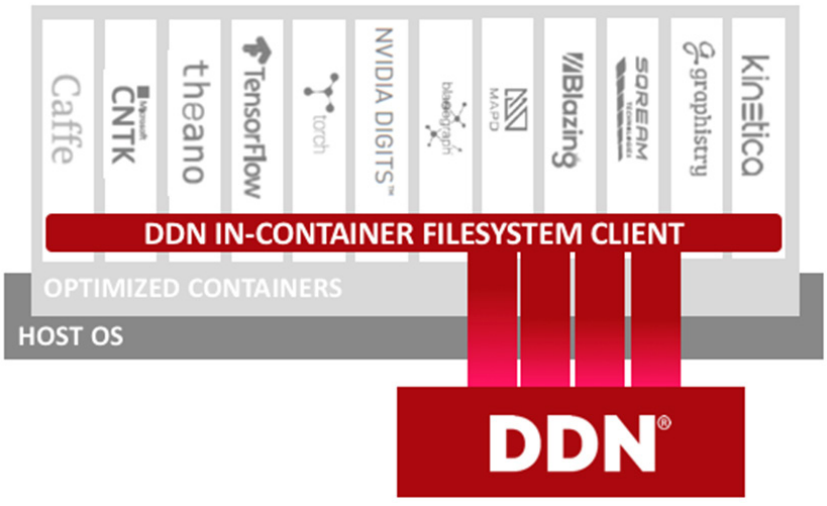
DDN EXAScaler Storage is now a new service in Google Cloud Marketplace (GCP) that's suited for dynamic, pay-as-you-go applications, from rapid simulation and prototyping to cloud bursting peak HPC workloads. AI and analytics in the cloud can now benefit from EXAScaler performance and capabilities.

A³I Solutions for NVIDIA DGX: DGX-1 and DGX-2

DDN A³I solutions enable and accelerate data pipelines for AI and DL workflows of all scale running on DGX servers. They are designed to provide high performance and capacity through a tight integration between DDN and NVIDIA platforms. Every layer of hardware and software engaged in delivering and storing data is optimized for fast, responsive and reliable file access. The DDN architecture provides highly optimized parallel data paths that run from containerized AI and DL applications running on DGX servers.

> "With the goal of expanding the boundaries of science as we know it today, we are excited about the arrival of advanced new technology that can dramatically increase the performance at scale of our systems, and specifically of our new Top 10 supercomputer, Frontera. DDN's new EXA5 has the power to provide the best I/O performance our users have ever experienced, and greatly reduce the I/O bottlenecks in large scale computation. We believe that EXA5 will play a role in many of the groundbreaking discoveries scientists will make with Frontera."
>
> – Dan Stanzione, Executive Director, Texas Advanced Computing Center (TACC)

The following reference architectures are specifically designed for AI and DL. They illustrate tight integration of DDN and NVIDIA platforms. Every solution has been designed and optimized



for DGX servers, and thoroughly tested by DDN in close collaboration with NVIDIA:

- Configuration for a single DGX-1/DGX-2 server with AI200
- Configuration for a single DGX-1/DGX-2 server with AI400
- Configuration for four DGX-1 servers with AI200
- Configuration for four DGX-1 servers with AI400
- Configuration for nine DGX-1 servers (DGX-1 POD) with AI200
- Configuration for nine DGX-1 servers (DGX-1 POD) with AI400
- Configuration for three DGX-2 servers (DGX-2 POD) with AI200
- Configuration for three DGX-2 servers (DGX-2 POD) with AI400

In terms of performance, the DDN A³I architecture provides high-throughput and low-latency data delivery for DL frameworks on the DGX-1 server. Extensive interoperability and performance testing has been completed using popular DL frameworks, notably TensorFlow, Horovod, Torch, PyTorch, NVIDIA TensorRT, Caffe, Caffe2, CNTK, MXNET, Theano. The figure below illustrates the GPU utilization and read activity from the AI200 storage appliance. The GPUs achieve maximum utilization, and the AI200 storage appliance delivers a steady stream of data to the application during the training process.

DDN is continuously engaged in developing technologies which enable breakthrough innovation and research in a broad range of use cases. Customers around the world are using DDN systems to expand their business, deliver innovation, and bring about social change that benefits everyone. The list below includes compelling use cases involving A³I solutions coupled with DGX servers to maximize the value of customer data and easily and reliably accelerate time to insight:

- AI powered retail checkout platform for video capture, processing and retention by Standard Cognition
- Real-time microscopy workflows by Microvolution
- AI platform for autonomous vehicle development by Uber
- Autonomous vehicle development by US Army Research Lab
- Life sciences by Siemens Healthineers
- Clinical Trials by National Institutes of Health (NIH)

GPU utilization and AI200 read throughput during Inception V³ training

Frameworks for AI and DL Workflows

DDN storage solutions incorporate three critical core technologies which come together in order to provide easy to manage, highly efficient access to AI applications and workflows. These technologies are: (i) a high performance parallelized filesystem storage appliance, (ii) a highly scalable, very low latency RDMA capable networking fabric, and (iii) comprehensive management, monitoring and reliability enhancing software.

With these technologies in place, the summary below highlights the acceleration benefits along with important benchmarks delivered by the DDN shared parallel architecture for widely-used AI and DL frameworks and convolutional neural networks:

TensorFlow

A commonly used open source framework for DL applications. Used for image, voice and sound recognition, TensorFlow applications depend on a large and diverse data set with rich media content. DDN systems provide the capacity needed to store and deliver massive heterogeneous data sets. DDN storage systems offer enhanced training throughput of 2x more images and 2x faster training times with TensorFlow for Inception v3.

Horovod

An open-source distributed DL framework for TensorFlow. The shared architecture of DDN systems provides a significant performance boost to saturate multiple GPUs engaged through distributed computing. This furthers the benefits of the Horovod DL framework for TensorFlow applications.

TensorRT

A high-performance DL inference optimizer from NVIDIA. DDN storage platforms enable TensorRT to deliver maximum improvements to neural networks using distributed computing at large scale.

Torch and PyTorch

Torch is an open-source scientific computing framework which provides a wide range of algorithms for DL that are optimized for parallel execution on GPUs. Pytorch is a python package based on Torch designed for rapid neural network development though an intuitive interface. DDN systems enhance and accelerate Torch and PyTorch frameworks. The DDN A³I solution's shared filesystem architecture provides accelerated distributed computing on multiple systems, with no data management overhead. Concurrent access to multiple data sets from all computing systems enables workflow flexibility, allowing complete freedom for data scientists to design and engage neural network training activities. DDN storage systems offer 3x faster training time for ResNet-152, 3x faster training time for VGG16, 3x faster training time for AlexNet, and 3x faster training time for ResNet-50 with PyTorch.

Caffe and Caffe2

Caffe, the convolutional architecture for fast feature embedding, is an open source DL framework that's optimized for image classification and image segmentation. DDN storage systems offer training throughput with 2.4x more images and 2x faster training time for Caffe GoogLeNet. Caffe2 is a flexible DL framework that extends the capabilities of the original Caffe and addresses its architectural bottlenecks. DDN storage systems offer 3x faster training time for AlexNet, and 2x faster training time for Inception v3 with Caffe2.

CNTK

The Microsoft Cognitive Toolkit is a DL framework highly optimized for speed, scale and accuracy. DDN storage systems offer enhanced training throughput of 3x more images and 2.5x faster training times with CNTK for ResNet-50.

MXNet

An open-source DL framework for training and deploying state of the art models, including deep neural networks, convolutional neural networks, and long short-term memory networks (LSTM). DDN storage systems offer enhanced training throughput of 2x more images for CIFAR-10, and 2.2x more images for Inception v3 with MXNet.

Theano

A python library for rapid efficient definition, optimization and evaluation of mathematical expressions using multi-dimensional arrays. DDN storage systems offer enhanced training throughput of 3.5x more images and 3x faster training times with Theano for AlexNet.

Partners Important Role for Leading-Edge Case Studies

DDN's strong channel partner solutions and programs have helped expand its presence in enterprise IT which has led to many compelling AI and DL use cases. Through the PartnerLink program, resellers possess all the tools they need to offer unique products and value-added service offerings that drive revenue growth, frequently in the most challenging market segments. Here is a short list of case studies across a wide spectrum of industries.

Robovision

Robovision, and its self-service deep learning platform, enables organizations to bring concrete, maintainable AI applications live in mere weeks. Partnering with Robovision allows common customers to accelerate the deployment and power of their AI applications.

XXII

French technology firm XXII provides computer vision solutions for solving problems in retail, security, manufacturing, insurance. The company's "frictionless retail" technology employs concepts such as autonomous check-out and real-time analytics to improve the shopping experience by creating seamless, in-store interactions to enable real-time customer service. XXII, a leading provider of real-time video analysis solutions, uses proprietary computer vision algorithms to allow customers to shop in stores without any friction and without having to wait in lines.

Max Delbrück Center

The Max Delbrück Center (MDC) for Molecular Medicine, in conjunction with Zuse Institute Berlin (ZIB), developed an IO intensive, high-throughput DL application using a large multi-terabyte imaging data set. The software saw a seamless upgrade from NVIDIA the DGX-1 to DGX-2 platform, and featured full GPU saturation with DGX-1 and DGX-2. MDC is a leading molecular biology and clinical research center in Germany, and ZIB is an interdisciplinary research institute for applied mathematics and data-intensive high-performance computing. MDC originally partnered with DDN in 2018 to greatly accelerate its applications and to efficiently manage large volumes of data. Using an A^{3}I architecture with an NVIDIA DGX-1 system and an AI200 from DDN, MDC was able to improve accuracy and precision within analyzed images while simultaneously accelerating training performance by 240 percent. More recent testing with an A^{3}I architecture comprised of NVIDIA's DGX-2 system and DDN's AI400 revealed results that more than doubled previous improvements.

Summary

Optimized storage solutions for AI and DL training and inference are critically important to data scientists, data engineers, and academic researchers, enabling these professionals to focus their complete attention on what really matters most – transforming valuable data assets into important insights with unmatched speed and accuracy.

In this technology guide, we've reviewed how optimized storage solves AI challenges, along with a detailed look at how DDN's A³I - Accelerated, Any-scale AI Solutions solve the hardened requirements of storage solutions required for workflows involving AI and DL. We also took a look at some popular frameworks for AI and DL workflows, as well as a number of compelling use cases for how optimized storage solutions have come to the rescue for some very important problems across a wide variety of industries. Here are some important takeaways when considering next steps to take in choosing your optimized storage solution:

1. Performance is a critical aspect of data storage for AI and DL workloads. Parallel data access is the key for keeping pace with the demands of these popular technologies.
2. Flexibility in the AI workflow is also vital in order to be able to deal with multiple data types, and engage multiple workflows.
3. Scalability allows you to plan ahead. Your needs today may be of limited scale. You may have small data needs today, but tomorrow it's likely that you'll be on quite a different path of collecting orders of magnitude more data as results of larger data sets, more sensors, and high bandwidth connectivity like 5G. With facilitating technologies like GPUs, suddenly you're able to collect and process more information.

The timing is critical for making strategic decisions about optimized storage solutions for managing increasing demands put in place by AI and DL applications. The hyper-competitive landscape is pervasive in many industries means that your competitors are making the same decisions to gain strategic advantage in the marketplace. To take key next steps for determining how you can facilitate important innovation by easily leveraging the power of new turnkey AI solutions for the data center be sure to visit DDN.

You may have small data needs today, but tomorrow it's likely that you'll be on quite a different path of collecting orders of magnitude more data as results of larger data sets, more sensors, and high bandwidth connectivity like 5G.

Expediting deployment and delivering acceleration in time-to-insight, DDN's pioneering approach enables you to manage the entire AI lifecycle and streamline your data center. DDN can demonstrate how their optimized storage solutions yield the following advantages:

- Easy to deploy AI solutions that immediately transform your AI concepts into business innovation
- Provide long-standing advantages that enable you to achieve high-performance AI at every stage of your organization's growth
- Show you how to realize the greatest technical and economic benefits through leveraging deep AI-expertise