WHITE PAPER

Pure Storage for Telco AI

How Pure Storage helps telcos deliver on the promises of AI

Contents

Telco operators are advancing toward AI-based autonomous networks to reduce costs and enable innovative digital services. Traditional storage systems, initially cost-focused and siloed, are being replaced as AI, machine learning, and deep learning drive new use cases like anomaly detection, self-healing, and traffic prediction. Pure Storage® delivers the unique data handling characteristics required to power telco AI use cases.

The Journey to AI in Telco Networks

Network automation and the journey to autonomous networks is a key objective to change the cost structure of telecom operators and provide the capacity to create and launch innovative digital services. Until recently, telco operators were using subscriber and network data for a variety of operational purposes such as network health monitoring, which includes alarming, demand and traffic analysis, congestion management, and more.

The storage architecture enabling these services relied on embedded or dedicated attached storage appliances, usually deployed on spinning disk hard drives or more recently on solid state drives (SSDs). The key factors for selecting these solutions revolved around storage capacity and cost. Data storage was considered by many as a commodity, with little differentiation between solutions resulting in a cost-driven, decision-making process. Most data storage solutions were deployed and dedicated to their respective network functions.

The recent progress in machine learning, deep learning and artificial intelligence (AI) has allowed telecoms to expand these use cases to automated pattern and anomaly detection, root cause analysis, self-healing, traffic prediction and end-to-end automation. These ambitious use cases require real-time network observability, coupled with powerful algorithms. They will allow the operator to react in real time to changes in network and traffic conditions, as well as to anticipate them and proactively reconfigure and optimize network elements.

Real time network observability and optimization require the capacity to extract the traffic, subscriber and network data from the network functions as fast as possible, so that they can be scrubbed and stored, ready for algorithmic processing.

A small part of the computing necessary for the detection of patterns and rules enforcement can take place at the edge (inference), close to the network functions, while the majority (model learning) must be processed centrally, in a private cloud in order to detect system-wide and network-wide patterns and to efficiently allocate resources for network and traffic optimization.

For instance, generative AI, through the processing of large language models (LLMs), can translate human intents and instructions into network rules and policies. The large amount of computing necessary for training the base models is why they are frequently implemented in public clouds, for economies of scale and an as-a-service consumption model. Retrieval augmented generation (RAG) techniques also allow extensions of LLM base models to be deployed on premises or in the private cloud.

Network infrastructure and architecture need to evolve to take full benefit of AI and automation. Many network operators start with solving the compute issue by sourcing GPUs, only to find that networking needs to be increased by the addition of SmartNICs and specialized accelerator cards. At that point, they realize that their traditional storage technology has become a bottleneck and needs to evolve as well.

To extract data from the network functions in real time and to manage and share them in a holistic fashion, the traditional dedicated disk-based appliances and direct attached storage can no longer satisfy the necessities of AI for telco networks.

Why Are Traditional Storage Architectures Not Suitable for Telco AI?

Data needs to be shared and read by many network functions simultaneously while it's being processed. Traditional architectures store data individually by network functions, then export to larger databases, then amalgamate into data lakes for processing. The process is lengthy, error-prone and negates the capacity to act or react in real time.

The data sets are increasingly varied, between large and small objects, data streams and files, and random and sequential read and write requirements. Legacy storage solutions require different systems for different use cases and data sets. This lengthens the data amalgamation necessary for automation at scale.

Data needs to be properly labeled, without limitations on metadata, annotation, and tags for both billions of small objects (e.g., event records), or very large ones (e.g., video files). Traditional storage solutions are designed either for small or large objects and struggle to accommodate both in the same architecture. They also have limitations in the amount of metadata per object. This increases cost and time to insight while reducing their capacity to evolve.

Data sets are live structures. They often exist in different formats and versions for different users. Traditional architectures are not able to handle multiple formats simultaneously, and versions of the same data sets require separate storage elements. This leads to data inconsistencies, corruption, and divergence of insight.

Performance is key in AI systems, and it's multidimensional. Storage solutions need to be able to simultaneously accommodate high throughput, scale out capacity and low latency. Traditional storage systems are built for capacity but not designed for high throughput and low latency, which dramatically reduces the performance of data pipelines.

Hybrid and multicloud become a key requirement for AI, as data needs to be exposed to access, transport, core, and OSS/BSS domains at the edge, and the private cloud and the public cloud simultaneously. Traditional storage solutions necessitate adaptation, translation, duplication, and migration to be able to function across cloud boundaries, which significantly increases their cost, while limiting their performance and capabilities.

As we have seen, the data storage architecture for a telecom network becomes a strategic infrastructure decision, and traditional storage solutions cannot accommodate AI and network automation at scale.

Storage Requirements for Telco AI

Perhaps the most important attribute for AI project storage is agility—the ability to grow from a few hundred gigabytes to petabytes, to perform well with rapidly changing mixed workloads, to serve data to training and production clients simultaneously throughout a project's life, and to support the data models used by project tools.

The attributes of an ideal AI storage solution are:

Performance Agility

- I/O performance that scales with capacity
- Rapid manipulation of billions of items, such as for randomization during training

Capacity Flexibility

- Wide capacity range (hundreds of gigabytes to petabytes) with easy, nondisruptive expansion
- High performance with billions of data items
- Range of cost points optimized for both active and seldom accessed data

Availability and Data Durability

- Continuous operation over decade-long project lifetimes
- Protection of data against loss due to hardware, software, and operational faults
- Nondisruptive hardware and software upgrade a nd replacement
- Seamless data sharing by development, training, and production

Space and Power Efficiency

- Low space and power requirements that free data center resources for power-hungry computation

Data Models

- Support for block, file, and object data models and common network protocols

Security

- Strong administrative authentication
- "Data at rest" encryption
- Protection against malware (especially ransomware) attacks

Operational Simplicity

- Nondisruptive modernization for continuous long-term productivity
- Support for AI projects' most-used interconnects and protocols
- Autonomous configuration (e.g., device groups, data placement, and protection).
- Self-tuning to adjust to rapidly changing mixed random/sequential I/O loads

Native Hybrid and Multicloud

- Data agility to cross cloud boundaries
- Centralized data lifecycle management
- Decide which data set is stored and where it is processed
- From edge for inference, to private cloud for optimization and automation, to public cloud for model training and replication

Pure Storage AI Products Portfolio

Portworx: The Pure Storage Secret Weapon

Portworx® by Pure Storage is a Kubernetes data services platform that provides persistent storage, data sharing and protection, workflow automation, and optional disaster recovery for containerized applications.

Portworx accelerates development of IT environments for the containerized applications used in most AI projects. Its software-defined storage model enables infrastructure-neutral access to data. Portworx supports any type of block storage, whether located on-premises or in a public or private cloud.

Portworx presents standardized virtual block or file storage devices to applications, regardless of the on premises or cloud technology used to instantiate it. It does this by making architect-defined storage classes available to developers. Storage classes standardize storage properties, simplifying self-service job creation and promoting stable, reliable development and production environments. In addition, Portworx includes templates that assist with setup for applications like Apache Kafka, Zookeeper, Elasticsearch, as well as for popular databases including SQL Server, MongoDB, Postgres, and Cassandra—all of which are commonly used in AI projects. It provides consistent development environments while enabling self-service job creation by data scientists and other developers.

The software-defined storage model enables Portworx to share data among multiple Kubernetes pods running separate jobs. This makes it particularly useful for model training, where running many concurrent jobs that share the same input data is key to rapid implementation. As an example, Figure 1 shows how Portworx can simplify deployment of a training application that utilizes data from a database.

Finally, Portworx provides fault tolerance by replicating its virtual storage devices to either on-premises resources or to a public cloud. It can also protect entire project environments by copying them to S3 objects in a public or on-premises private cloud.

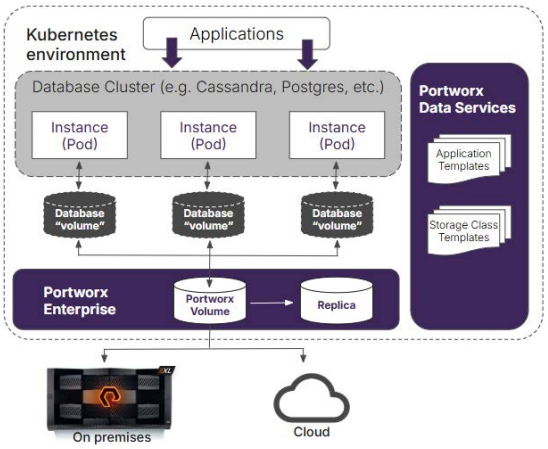

FIGURE 1 Using Portworx with a database app.

Pure Storage Products Portfolio

The Pure Storage portfolio of all-flash storage systems, illustrated in Figure 2, includes three FlashArray™ scale-up Unified Block and File (UBF) appliances, a FlashBlade® scale-out Unified Fast File and Object (UFFO) appliance, and two Unified Data Repository (UDR) appliances, one based on FlashArray and the other on FlashBlade, for low-cost, large-scale storage. All systems support broad ranges of capacity that are easily expandable while online. Each is optimized for specific capacity ranges, data type(s), and cost/performance targets. With these devices, Pure Storage can satisfy virtually any AI storage requirement from project conception through model training, and on into production.

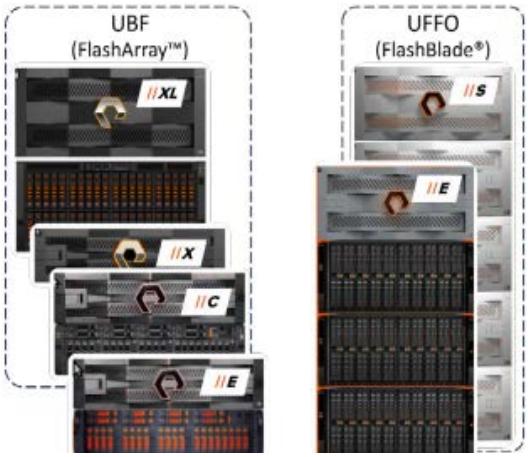

FIGURE 2 The Pure Storage portfolio for AI

Features Common to All Systems

All Pure Storage systems share key properties:

Reliability and Availability

- Systems are designed to continue operating when any internal component fails. For example, they survive at least two (in most cases more) rare DirectFlash® Module (DFM) failures that overlap in time without loss of data or client access to it.

Efficiency

- Systems optimize capacity utilization by thin provisioning, deferring space allocation until clients write data. They allocate space autonomously to balance utilization and load. The high density of DFMs minimizes system "footprint," power consumption, and ultimately, e-waste.

Simplicity

- Systems minimize administrative tasks to the greatest extent possible. There are no "device groups" to manage, and no data placement or protection decisions to make. Systems report status and events to the Pure1® cloud frequently. Pure1 analyzes behavior and proactively initiates any necessary service operations.

Continuity and Longevity

- Systems are designed for lifetimes of a decade or more of continuous operation with no planned downtime or service outages, even during software and hardware upgrades and modernizations.

Evergreen Subscriptions

- Evergreen® subscriptions that include regular software and hardware updates and periodic modernizations are perhaps the most important benefit for AI projects of long duration. Pure Storage offers subscriptions both for purchased systems and for storage delivered as a managed service (Evergreen//One™). Technical briefs TB-230601f and TB-230601o, available at https://support.purestorage.com or from a Pure Storage representative, describe the company's Evergreen offerings in more detail.

Product-specific Features

Spectrum of Performance and Cost Options

- From latency-optimized FlashArray//XL™ for rapid response in production to throughput-optimized FlashBlade//S™ for training with very large data sets, to cost-optimized FlashArray//E™ and FlashBlade//E™ for less-active data (e.g., raw data, feature stores, etc.), the Pure Storage product line offers cost/performance options that span AI project requirements. Pure Storage systems' very high-performing metadata operations on large numbers of files and objects make them particularly suitable for AI project training.

Capacity Flexibility

- With maximum capacities ranging from the FlashArray 918TB (effective) to FlashBlade//S nearly 20PB (physical), Pure Storage systems can support many in-house AI projects from concept through production with a single project-wide data hub. Systems can "start small" in a single chassis with minimal physical capacity and be expanded up to a model's maximum supported capacity without interrupting service to applications. Where multiple systems are required, for capacity, performance, cost, or data model reasons, Pure1 centralizes storage management and supports AI-based "what if" capacity planning for the users' entire "fleet" of Pure Storage systems.

Data Reduction

- Pure Storage systems optimize flash utilization by compressing data prior to storing it. In addition, FlashArray systems achieve further efficiency by deduplicating blocks of data, replacing duplicate blocks with links to a single stored instance. Deduplication works well with structured data (e.g., databases, tables, and more).

NVIDIA Collaboration

- With approximately 80% market share, NVIDIA Corporation is the acknowledged leader in AI computation. Since 2017, Pure Storage has collaborated with NVIDIA to develop solutions for AI. The collaboration has resulted in multiple solutions:
  - FlashBlade//S is a certified Ethernet-based solution with NVIDIA DGX SuperPOD. This includes a full reference architecture (log-in required).
  - A jointly-developed AIRI® Pure Storage NVIDIA DGX BasePOD Reference Architecture for AI, based on the NVIDIA DGX servers and network fabric, coupled with the FlashBlade//S storage systems. As a NVIDIA BasePOD certified reference architecture, AIRI eliminates the design, deployment, and management complexity inherent in custom-crafted AI infrastructures.
  - A joint solution between NVIDIA, Cisco, Red Hat and Pure Storage: FlashStack® for Generative AI Inferencing. See the Design Guide.

- The Pure Storage and NVIDIA collaboration has also resulted in:
  - The implementation of the NVIDIA GPUDirect storage protocol to transfer data directly between FlashBlade//S storage and NVIDIA GPUs, bypassing control CPU memory.
  - Storage partner validation for FlashBlade//S in NVIDIA-certified OVX L40S reference architectures offered by major server vendors. When combined with FlashBlade//S storage, OVX-certified servers are complete AI platforms that accelerate small model training and fine-tuning, as well as GenAI RAG and production inference workloads.
  - Finally, when used in conjunction with Portworx, the NVIDIA device plugin for Kubernetes provides comprehensive management and scheduling of both GPU and storage resources at all AI project stages.

Conclusion

The Pure Storage system portfolio includes storage for all phases of AI projects, large and small. Pure systems relieve IT, data scientists, and MLOps teams from most common storage management tasks. With them, data scientists can concentrate on modeling and MLOps teams can provide reliable, scalable, high-performing environments for project data with shareable storage that expands to meet both training and production needs without disruption.

Available in performance-optimized and capacity-optimized models that scale on demand, the Pure Storage platform accelerates and enhances telco AI projects for network, energy and spectrum optimization, Open and Virtualized RAN Intelligence Controller and Service Management and Orchestration, and many other use cases. It can do so while sharing storage capacity and I/O bandwidth with other data-intensive applications such as analytics, database backup and restore, software development, media and entertainment post-production, electronic design automation (EDA) and more.

Portworx takes the guesswork out of creating robust, scalable Kubernetes environments for containerized AI training and production applications. Its templates simplify configuring and implementing applications and databases commonly used in AI projects; its built-in storage services enable data sharing and backup of both project data and entire project environments.

PS2726-01-en 02/25