International Journal of Computer Trends and Technology
ISSN: 2231–2803 / https://doi.org/10.14445/22312803/IJCTT-V73I5P116
Volume 73 Issue 5, 125-132, May 2025

Original Article

Solid State Drive: Opportunity, Method, and Apparatus to Address Artificial Intelligence Infrastructure Data Storage Challenges

Joydeep Das

Senior Technical Program/Product Manager, Pure Storage Inc. California, USA.

[1]Corresponding Author : joydeep1977@gmail.com

Received: 30 March 2025 Revised: 02 May 2025 Accepted: 16 May 2025 Published: 31 May 2025

Abstract *- This document examines the computing and data storage infrastructure requirements necessary to support the ever-evolving demands of artificial intelligence workloads. It deeply explores AI trends and evaluates whether AI infrastructure presents a genuine opportunity for solid-state storage technology to broaden its use cases. This document discusses the history of AI, the diverse workload requirements for data storage across various phases of AI infrastructure, and the risks associated with current artificial infrastructure demands. Additionally, it highlights the potential for solid-state drives to deliver more efficient data storage features for future AI infrastructure deployment.*

Keywords *- Artificial Intelligence, Data storage, Solid state drives, Infrastructure, Computational storage, CXL Memory, Inference, Compute offloading, CXL hybrid module, NVMe, NVMe-OF.*

1. Introduction

The resurgence of Artificial Intelligence, particularly generative AI, has resulted in a substantial increase in investments, research, and the development of new technological advancements in computational power and data storage technologies. While investments in computing and data storage infrastructure are growing exponentially, not all stages of artificial infrastructure require the same or similar data storage workloads. This discrepancy means that the type of storage capabilities varies across different workflow stages. This document explores various data storage workloads, capabilities, and future growth opportunities for solid-state storage technologies throughout the stages of AI infrastructure. It also examines the challenges of AI infrastructure in achieving artificial general intelligence with the current infrastructure approach.

2. Artificial Intelligence – Brief History and Composition of AI Infrastructure

2.1. History of AI

Contrary to popular belief, artificial intelligence began in the 1950s with Alan Turing and the Dartmouth workshop. The 1960s saw a significant leap with MIT's introduction of the first natural language processing capability through ELIZA. However, progress stalled between the mid-1960s and late 1970s due to a lack of computational power, a period referred to as "AI Winter." The 1980s and beyond experienced a resurgence in AI with the development of neural network technologies, supervised and unsupervised learning algorithms, and deep learning architecture. The 1990s marked the introduction of "Deep Blue," the first AI supercomputer launched by IBM. The current generation of the artificial technology boom started in the early 2000s with the rise of big data analytics, machine learning, and more sophisticated deep learning algorithms, leading to the rapid growth of AI and the emergence of generative AI starting in the 2020s.

2.2. Composition of AI Infrastructure

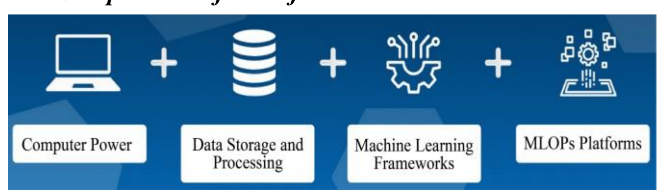

Fig. 1 Composition of High-level AI Infrastructure

The entire artificial intelligence infrastructure is standing on four pillars.

- Computer processing power - GPU (Parallel Processing) and TPU (Deep Learning)
- Data storage and processing – Cloud/Hybrid/On-prem scale-out storage (File, Object)
- Machine learning frameworks – TensorFlow, PyTorch, SageMaker, Azure Machine Learning
- MLOps platforms - ML Flow, Kube Flow, SageMaker, and many others

3. Materials and Methods

AI's entire success and lifecycle depend on the availability of clean, reliable, and accessible data. So, data for AI is == Oxygen for the human brain/body. Let us break it down and see where scalable data storage comes from and how it plays a vital role in different stages of AI infrastructure.

3.1. Workload

The workload is broken down into four stages.

3.1.1. Ingestion

Collection of data into infrastructure

3.1.2. Data Preparation

Cleaning, preparing data for training

3.1.3. Training

Using the foundational model (i.e., GPT, Llama, Gemini, etc.)

3.1.4. Inference

Accessing trained data on the data storage and running decisions based on training data.

3.2. Availability

The success of data availability depends on three following factors

3.2.1. High Performance

These deals with how fast data can be stored and retrieved to/from data storage media.

3.2.2. Storage Management

This part determines how efficiently infrastructure manages several hundred to thousands of data storage hardware nodes in a cluster or multiple clusters and allows fault tolerance.

3.2.3. Integration

Integrating a storage cluster into the compute infrastructure is vital to data management and availability.

3.3. Efficiency

The efficiency of the data storage subsystems depends on the following three capabilities of storage technology.

3.3.1. De-Duplication

Data de-duplication is a data storage algorithm that allows a host system to identify the same data in a data stream and store only one instance of the data in storage media. This allows data storage to minimize the number of writes into the storage media and allows the host system to retrieve data faster.

3.3.2. Compression

Compression is a technology in which the host system compresses all the data into a smaller size. This allows the host system to transfer a limited amount of data into media, which enables the underlying storage subsystem to store and retrieve data faster.

3.3.3. Tiering

Tired storage is a mechanism where data is stored in different storage media (solid state, mechanical hard drives, tape drives) based on their access pattern. Typically, data are segregated into hot, warm, and cold data, where hot data is most frequently accessed and stored in a solid-state drive, system memory, nonvolatile RAM, etc.

Warm data are accessed less frequently and stored in solid-state and mechanical hard drives. Cold data are primarily for archival purposes and stored in mechanical hard drives or tape drives.

3.3.4. Data Locality

Data locality typically deals with how fast the host computer memory can access the data. Data stored closer to system memory is accessed faster than data stored in a far location where host systems need additional mechanisms like network bandwidth and PCIe expansion hardware to retrieve the data.

3.4. Data Protection

Data protection is another aspect of the storage infrastructure, where storage subsystems must store data for future access and legal requirements. These are typically divided into three categories.

3.4.1. Security

Securing stored data is not only critical for the infrastructure to run efficiently but also critical for any analyzing engine, like an AI foundational model, to run properly to train its decision-making algorithm

3.4.2. Backup and Restore

Backup and restore deals with disaster recovery data corruption prevention and facilitates data retention

3.4.3. Retention

Data retention is governed by local, national, and international laws and is critical to data storage management systems.

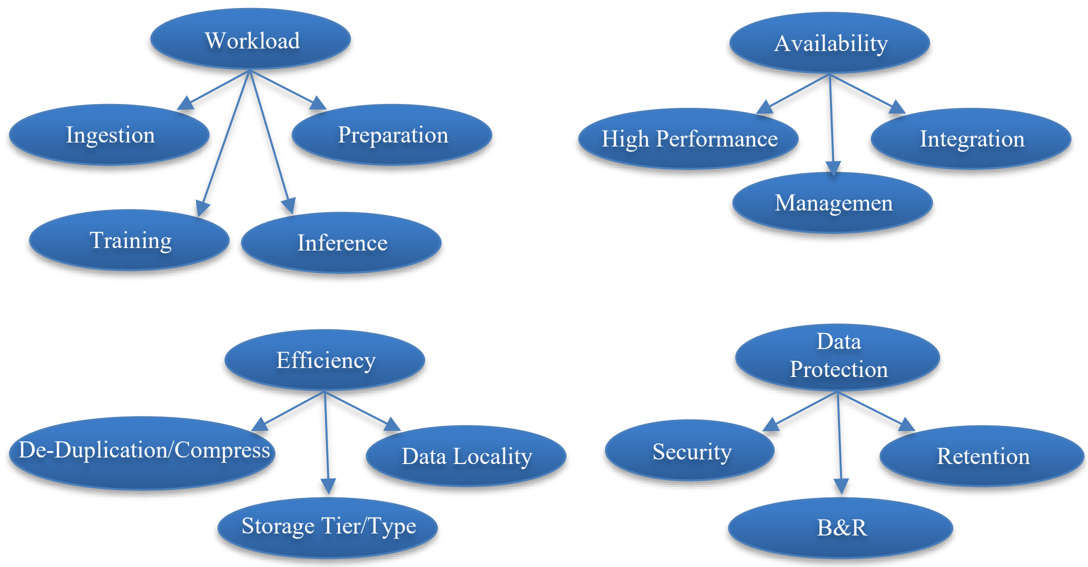

Fig. 2 Component of scalable storage Infrastructure

4. Results and Discussion

4.1. Where AI is Heading – Business Case Study

Before proceeding further into the data storage requirements for AI infrastructure, it is important to understand the overall AI investment landscape and why solid-state drive manufacturers should invest in the future of AI.

- The global AI market is at $600B and expected to reach $1.8T by 2030
- US market projected to $300B by 2026
- AI market CAGR is to the tune of 37% since 2023 and is expected to remain strong.
- 48% of businesses across all industries use some form of AI today
- AI software revenue grew 12x since 2018($10B in 2018 to $126B in 2025)
- AI infrastructure investment (server and storage) is expected to grow 44x ($5B in 2023 to $220B in 2028)
- 22% annual capacity growth projected between 2025-2030

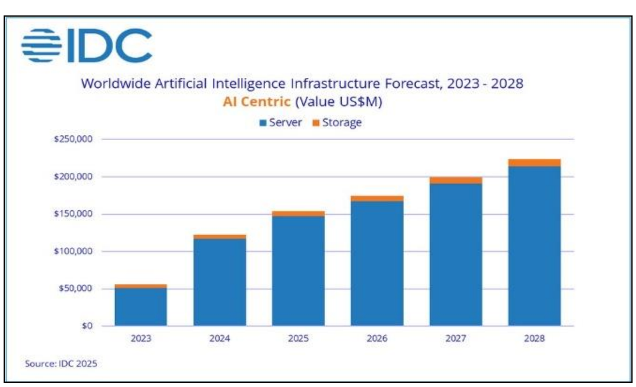

Fig. 3 IDC forecast for AI infrastructure investment

4.2. Opportunity for Solid State Drive Technology

Advancements and rapid progress in artificial intelligence presented a unique opportunity for solid-state drive technology to address several data storage challenges the AI infrastructure is encountering, along with replacing slower and more power-consuming mechanical hard disk drives. We will break down different stages of data storage requirements to run AI infrastructure, but let us look into the three big \ pillars where the advancement of solid-state drive technology would bring maximum impact.

4.2.1. Speed and Efficiency

Improving the speed and efficiency of solid-state drives will be needed to make them most relevant to AI storage infrastructure. Solid-state drive manufacturers have the opportunity to implement this in three broad areas.

Faster Adoption of PCIe Standard

Currently, most data storage platforms use PCIe gen-4 or gen-5. However, the PCIe gen-6 and beyond standard is already available in the industry, and the quicker adoption rate of the latest generation of PCIe standard will bring significant speed increases to solid-state drives, especially for solid-state NVMe drives.

Cost and Capacity Optimization

Higher capacity, better speed, and reduced cost (through QLC NAND adoption) would bring a significant advantage to solid-state drives, whereas replacing mechanical hard drives with solid-state drives would bring lower TCO and lower data centre power budgeting advantage. Disintegrating FTL (Flash Translation Layer) from drive firmware to host (open channel SSD) or similar technology would allow solid-state drive manufacturers to eliminate the need for a proportionate increase of DRAM (Dynamic random-access memory) capacity as they try to increase the drive capacity by packing more NAND die packs into PCBA. Adopting standard form factors like different EDSSF standards (E3.S, E2.S) would allow for more NAND packs without increasing the power requirements of the solid-state drive. A combination of QLC NAND, FTL disintegration, and adoption of EDSSF standards will allow solid-state drives to be more technically and business-attractive for hyperscale AI infrastructure providers.

Bringing Compute Power where Data

Bringing Compute Power where Data is residing generally refers to in-situ processing. Before we explore in-situ processing, it is important to understand how data is currently being processed in a computer server environment.

MMIO (Memory Mapped IO)

When a CPU instruction reads the memory of a device's MMIO region, a Memory Read Request Transaction Layer Packet (MemRd TLP) is generated and transferred from the Root Complex of the host machine to the device. This type of TLP informs the receiver that the sender wishes to read a certain number of bytes from the receiver. This packet expects the device to respond with the contents at the requested address as soon as possible.

DMA (Direct Memory Access)

The first step in enabling DMA (Direct Memory Access) is for the device driver to request a memory buffer from the OS using an API call. This memory must typically be a contiguous block of physical memory, which can be challenging for the OS to allocate, especially on systems with limited resources. However, modern enhancements like Scatter-Gather and IOMMU also allow non-contiguous memory. Once allocated, the API returns a logical (often physical) address for the device to access via DMA. The driver then fills the buffer with the data (e.g., 01 02 03 04 pattern) intended for transfer to the device.

The second step in the DMA transfer is configuring the device with the necessary information to perform the transaction. This setup depends on the specific DMA interface of the device, which varies and requires referencing standards (like NVMe) or working with the hardware designer. For example, a simplified device interface uses a BAR0 MMIO region, meaning the driver must write configuration data to specific memory-mapped registers in BAR0. The driver knows these register locations and includes them in its code. Now, the OS driver will need to write the necessary values into the registers using the mapped memory of BAR0 for the device (how it mapped this memory depends on the OS). “Target Memory" determines the memory we want to copy from the device and maps to a region of memory in the device’s on-board RAM, which acts as destination memory. OS now creates a “DMA Buffer” that allocates a chunk of memory at a memory address and performs as the source address. At this point, the driver has configured all the registers necessary to perform the transfer. The last step is to write a value to the initiate transfer register, which acts as the “doorbell register” that begins the transfer. As soon as this value is written, the device will drive the DMA transfer and execute it independently of the driver or the CPU’s involvement.

The third part of DMA deals with the doorbell register. The device's DMA Engine handles the entire transfer process. It uses configuration data written to BAR0 registers by the OS driver to send Transaction Layer Packets (TLPs) over the PCIe link and perform the necessary memory transactions.

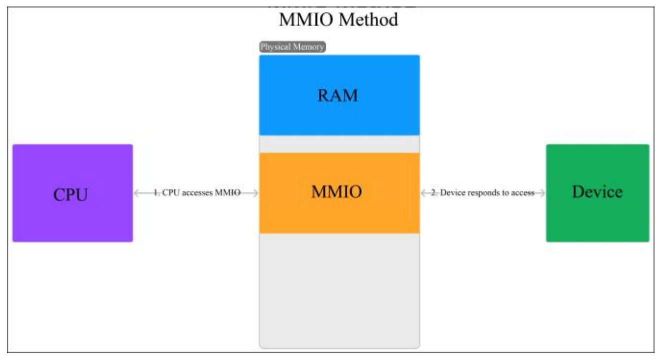

Fig. 4 Memory map IO process

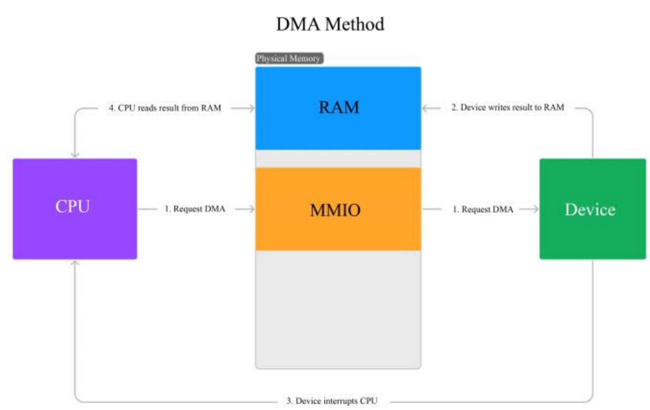

Fig. 5 Direct memory access

Now that we have a fair knowledge of DMA and MMIO, let us focus on the in-situ processing. As we have seen above, in both cases of MMIO and DMA, the host operating system is entirely dependent on the processing capacity of the system’s CPU and/or GPU and total available host memory to request data residing on storage media, process the request, and derive the resulting outcome. Current AI infrastructure uses this traditional architecture for ingesting, data preparation, training, and achieving the data. However, the inference workload entirely relies on GPU (graphical processing unit) or CPU (central processing unit) to run the entire inference engine, where GPU/CPU uses various host OS and data storage protocols to access and process the data from a storage device to arrive at a decision.

The existing model does not provide data storage with any ability to process inference data unless the data storage manufacturer adopts the technical advantage of computational storage and combines it with CXL (Compute Express Link) hybrid module where the inference workload can be processed inside the data storage device with its own compute capability and DRAM attached to CXL hybrid module.

Bringing computing power into storage devices would allow AI infrastructure to avoid over-reliance on GPU clusters or increasing GPU clusters proportionately as the more complex foundational model is being developed. In-situ computing capability for text, machine learning data, and natural language inference engines would provide solid-state drive manufacturers access to a vast and critical segment of overall artificial intelligence inference engine space for rapid opportunity growth.

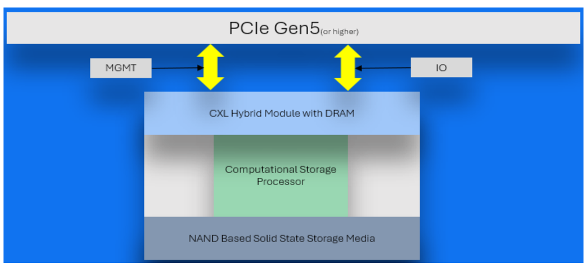

Fig. 6 Computational storage with CXL Hybrid

4.3. AI Infrastructure Storage Workload and Solid-State Drive Capability Matrix

As discussed under section 3.1, data storage requirements for successfully running artificial infrastructure can be broken down into six categories. Each category has its requirements and serves a specific purpose to achieve the outcome of the artificial intelligence model. In this section, let us break down the sections and identify how solid-state drives have the opportunity to address the storage infrastructure needs.

4.3.1. Data Ingestion

Data ingestion is the 1st stage of processing, where an artificial intelligence engine ingests various data sources from different sources where data are stored. In this stage, storage needs to have a high capacity with heavy read performance and moderate to light write performance

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt either E2.S or U.2 form factor, which will allow maximizing the storage to be capable of high capacity without increasing power budgeting of the drive with a combination of QLC NAND to achieve high-capacity storage with reasonable power budget. Host-based NVMe, NVMe-oF, and 12/24G SAS protocol-supported storage would suffice the workload requirement.

4.3.2. Data Preparation

Data preparation is the 2nd stage, where ingested data are cleaned and prepared in a structured orientation so all ingested data are prepared for the 3rd stage, where the AI foundational model could train all the data sets. This stage typically sees moderate capacity and heavy sequential read and write workloads.

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt either E2.S or U.2 form factor, which will allow maximizing the storage to be capable of moderate to high capacity without increasing power budgeting of the drive with a combination of QLC NAND to achieve the capacity point of the storage with reasonable power budget. To support heavy sequential read and write, improving the solid-state drive's channel speed and introducing a buffer chip into the NAND controller would allow support for heavy read and write performance while using QLC NAND on the storage device.

4.3.3. Training

Training is the 3rd stage, where artificial intelligence foundation models like Lalma, GPT, Gemini, etc., run training on all prepared data sources to analyse different data sources and create millions of checkpoints throughout this stage. This stage typically sees heavy random read and moderate write IO.

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt either E2.S or U.2 form factor, which will allow storage to be capable of moderate capacity without increasing power budgeting of the drive and combining TLC NAND to achieve a capacity point of the storage with speedy NAND type. A solid-state drive with hybrid NAND(SLC+QLC) would also bring a similar performance envelope while decreasing the cost of the storage drive due to the lower cost of QLC NAND. Adopting TLC NAND with an interface chip would allow improved performance to support heavy random read and moderate write, which would be necessary for the training workload.

4.3.4. Checkpointing

Checkpoint or data checkpointing-based training data is the 4th stage of data infrastructure where AI foundational models run training on the prepared data and, for every variance, create a checkpoint of the data set. The volume of the entire checkpoint for a moderately complex may run into a few million checkpoints being created for later use during inference. This stage typically sees high random read and high sequential write.

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt either E2.S or U.2 form factor, which will allow storage to be capable of moderate capacity without increasing power budgeting of the drive and combining TLC NAND to achieve a capacity point of the storage with high-speed NAND type. To support heavy Adopting TLC NAND with an interface chip with high read and random write would improve performance, which would be necessary for checkpoint workload.

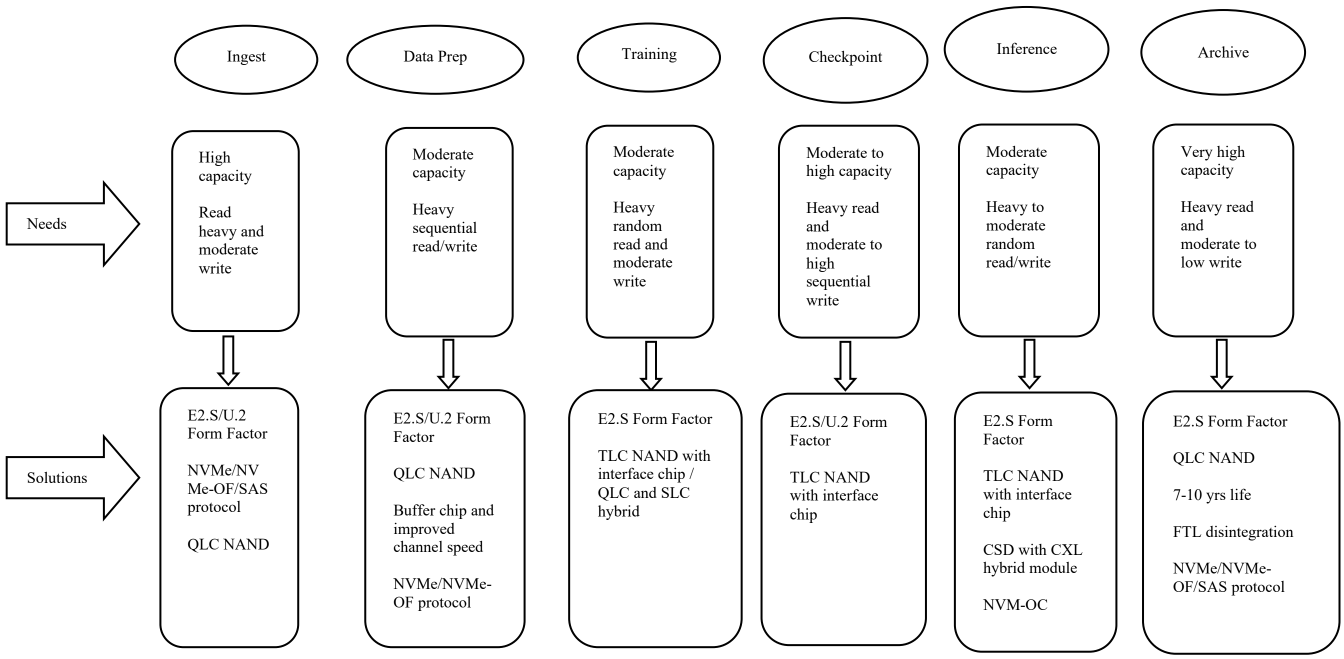

Fig. 7 AI Infrastructure storage workloads and SSD capability matrix

4.3.5. Inference

The Inference or decision stage is the 5th stage in the artificial intelligence data infrastructure, where inference engines typically use GPU clusters to access data storage across various storage nodes and run inference and decision-making processes based on the checkpoint's foundational models created during 4th stage. Data storage typically sees heavy random read and moderate write IO in this stage.

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt the E2.S form factor, which will allow storage to be capable of moderate capacity without increasing power budgeting of the drive and combining TLC NAND to achieve a capacity point of the storage with high-speed NAND type. However, the introduction of computational storage with the CXL hybrid module (discussed under section 4.2.1.3 for inference workload processing will offload compute load into storage where the storage device will be able to handle the majority of the inference directly in the storage device to improve significantly the performance while reducing the dependency on GPU for inference workload.

4.3.6. Archive

The archive stage is the 6th and last of the artificial intelligence data infrastructure where data collected from stage 1 to inference run on stage 5 is stored in long team storage where the host would be able to run workloads like big data analytics and machine learning algorithm to fine-tune future inference as well as creating synthetic data sets for various simulated decision situations. Data storage typically sees heavy reads and moderate to low write IO in this stage.

Solid-State Drive Capabilities

To address this stage of storage requirements, solid-state drives need to adopt either E2.S or U.2 form factor, which will allow maximizing the storage capacity without increasing the power budgeting of the drive with a combination of QLC NAND to achieve the highest capacity point of the storage.

However, mechanical hard disks and tape drives dominate this data storage segment. To get into this segment of data storage, solid-state drives not only need to offer the highest capacity possible but also move FTL (Flash Translation Layer) out of drive firmware to host or disintegrate FTL up to a certain extent (i.e., open channel SSD) to reduce the dependency of having additional DRAM as the capacity of the storage drive increases. Also, extending the shelf life of the solid-state drive up to 7-10 years would be required to gain acceptance into data archive workload space.

5. Conclusion

Recent advancements in artificial intelligence and the need for more and more infrastructure to fine-tune and advance artificial intelligence foundation models' development provide a unique opportunity for drive technologies to become mainstream into the entire spectrum of AI data storage infrastructure. However, it is critical to assess the pros and cons we have in the AI market to derive a conclusion. Let us do a deep dive and analyze the data.

Pros:

- Unlike conventional belief, the concept and advancement of artificial intelligence started more than 70 years ago. Less computing power prevented the advancement of artificial Intelligence, but GPU changed the dynamics drastically.
- There is possibly no company that exists today that does not have an AI strategy (small, big or bold) in place.
- Ever-improving foundational models push the boundary of AI capability daily, requiring ever-increasing performant infrastructure and thoughts beyond convention.
- SSD market to support AI Infrastructure is growing at 20% CAGR and is expected to remain strong
- The next 5-10 years of revenue growth will come to the companies working on
    1. AI Infrastructure
    2. Foundational Model
    3. Wrapper Agents

- Adopting newer technologies and promoting/developing smarter, less powerful, hungry, and larger SSD storage has the potential to become mainstream in all spectrums of the AI Infra and storage landscape.

Cons:

- The unavailability of alternate compute architecture and over-reliance on GPU would lead to an unsustainable power requirement to advance AI development and has the potential to bring back "AI Winter" again unless technological advancement progresses enough with quantum computing and neural processing units.
- The lack of usable use cases and slow ROI for enterprises to monetize AI investment potentially forces enterprises to rethink their AI investment strategy
- Potentially low ROI during the initial adoption phase of new SSD technology adoption due to introducing new features and technology standards without a proven enterprise success story. So, adoption and advancement in newer solid-state drive technology requires a multi-year view and leadership advocacy to see long-term benefits.
- The rapid advancement of synthetic DNA data storage in the last 5 years might pose a real challenge for current mainstream storage technologies due to DNA storage technologies' ability to store data over 10000 years and extremely low latency profile. However, this risk is not immediate since DNA storage is still under an experimental model and has not yet been implemented at the enterprise level.

Funding Statement

This research did not receive any special grants from any funding body. The author of this article supports the publication's relevant fee. The author performed the primary research without any sponsor's financial contributions from various organizations, commercial entities, or governmental agencies. The authors also take full responsibility for designing, conducting, and analyzing the research. Every claim and suggestion in this paper stems from the literature and a critical analysis of existing data governance frameworks.

Acknowledgments

The author acknowledges all the reviewers for their fruitful reviews and time.

References

[1] Dina Fakhry et al., "A Review on Computational Storage Devices and Near Memory Computing for High Performance Applications," *Memories - Materials, Devices, Circuits and Systems,* vol. 4, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Qirui Yang et al., "Performance Evaluation on CXL-enabled Hybrid Memory Pool," *2022 IEEE International Conference on Networking, Architecture and Storage (NAS),* 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] IDC Report, 2025. [Online]. Available: https://my.idc.com/getdoc.jsp?containerId=prUS52758624

[4] Storage Network Industry Association. [Online]. Available: https://www.snia.org/forums/cmsi/knowledge/formfactors#edsff

[5] Ion Stoica et al, "A Berkeley View of Systems Challenges for AI," *arXiv Preprint*, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[6] P.M.A Van Ooijen, Erfan Darzidehkalani, and Andre Dekker, "AI Technical Considerations: Data Storage, Cloud usage and AI Pipeline," *arXiv Preprint*, 2022, [CrossRef] [Google Scholar] [Publisher Link]

[7] Yang Lu, "Artificial Intelligence: A Survey on Evolution, Models, Applications and Future Trends," *Journal of Management Analytics*, vol. 6, no. 1, pp. 1–29. 2019. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sandeep Pandey, Snigdha Gupta, and Shubham Chhajed, "ROI of AI: Effectiveness and Measurement," *International Journal of Engineering Research & Technology (IJERT)*, vol. 10, no. 5, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Lee Organick et al, “An Empirical Comparison of Preservation Methods for Synthetic DNA Data Storage,” *Small Methods*, vol. 5, no. 5, 2021. [CrossRef] [Google Scholar] [Publisher Link]
[10] Sukhpal Singh Gill et al, “AI for Next Generation Computing: Emerging Trends and Future Directions,” *Internet of Things*, vol. 19, 2022. [CrossRef] [Google Scholar] [Publisher Link]