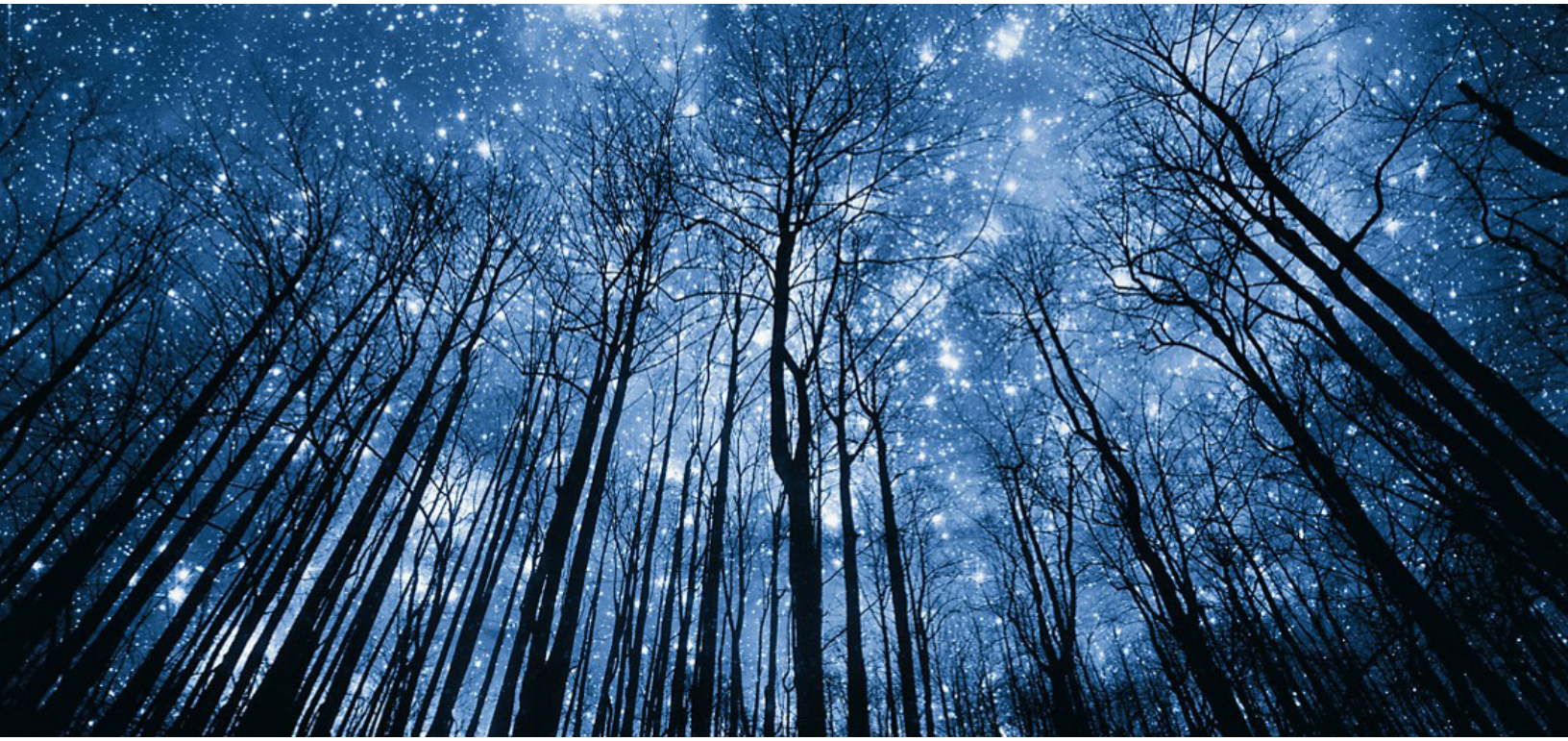
DELL Technologies

DNA: A NATURAL REALM FOR DATA STORAGE



Parleen Oberoi
Associate Sales Engineer Analyst
Dell Technologies

Keerthan S J

DELL Technologies
Proven Professional

Knowledge Sharing Article

The Dell Technologies Proven Professional Certification program validates a wide range of skills and competencies across multiple technologies and products.

From Associate, entry-level courses to Expert-level, experience-based exams, all professionals in or looking to begin a career in IT benefit from industry-leading training and certification paths from one of the world's most trusted technology partners.

Proven Professional certifications include:

- Cloud
- Converged and Hyperconverged Infrastructure
- Data Protection
- Data Science
- Networking
- Security
- Servers
- Storage

Courses are offered to meet different learning styles and schedules, including self-paced On Demand, remote-based Virtual Instructor-Led and in-person classrooms.

Whether you are an experienced IT professional or just getting started, Dell Technologies Proven Professional certifications are designed to clearly signal proficiency to colleagues and employers.

Learn more at www.dell.com/certification

TABLE OF CONTENTS

Introduction

The generation of digital data is rapidly increasing, surpassing current storage technologies' capacity. Traditional storage mediums such as magnetic tapes, hard drives, and solid-state drives have been used for decades but are facing limitations in terms of storage density and longevity.

As the amount of digital data continues to grow, the challenge of storing it cost-effectively and efficiently becomes increasingly important. DNA storage may prove to be an effective and unconventional solution for archiving copious amounts of information.

DNA storage presents a viable alternative as it is exceptionally stable, takes up minimal physical space, and can be read using technology already used for genetic analysis, ensuring its longevity.

Abstract

Data is very crucial for the evolution of humanity and therefore storing and managing it is an important requirement in every era. We started storing and managing information while we were cave people by carving patterns or drawing on the rocks and evolved all the way up to store data in the form of bits in storage drives. We still use several types of drives to store our beloved data, but the amount of data we generate currently 'drives' these storage mediums to their limits.

"Your Storage is full" This is the annoying notification we get when we run out of space. All we do is manually clear the least important data or back it up in the cloud; who is behind the scenes are these storage drives, again. The storage devices which we are currently relying on may no longer support our future data storage demands for storing the data. We generate roughly 2.5 quintillion bytes of data every single day and that is why this is truly a data era. This era is strongly demanding a revolution in the evolution of storage drives.

Storing all the data generated in this timeline and managing them is a Gordian Knot. But Is there a way to store all of these ever-growing data? The answer lies in the storage drives of Data center created by Nature: Deoxyribonucleic acid DNA.DNA carries genetic information of an organism and the constituent of this information is combinations of bases A,G,C and T. Hopping from storing data in the form of 'base' 2 to nucleobase is promising in bringing in the revolution which the data era needs. DNA not only has the ability to store data, but it can outmatch the current technology with its dense capabilities.

This whitepaper spreads light on the scope of storing the data in DNA, and the benefits which help us to solve this crucial problem of storing all the data and its future. There is a strong possibility that you might access this whitepaper to read from the DNA where it is stored.

Conventional and Modern Storage media

Storage media has evolved in an elevation with respect to the demanding needs of data users. “Conventional storage media” refers to the various types of physical storage devices that are used to store and retrieve digital data. These storage devices include: -

- **Hard Disk Drives (HDD)**: A hard disk drive (HDD) is a storage device that uses magnetic disks to store and retrieve data. An HDD comprises several components, including a spindle motor, a read/write head, and one or more disks (platters) coated in the magnetic material. Data is stored on an HDD in a series of concentric circles called tracks, divided into smaller sections called sectors. The read/write head, mounted on an arm that moves across the disk, reads data from, and writes to the disk. When the computer wants to read or write data, it sends a command to the HDD's controller, which interprets the command and moves the read/write head to the correct location on the disk. To write data to an HDD, the write head magnetizes a small area on the disk surface to store the data. To read data, the read head detects the magnetic fields on the disk and converts them into binary data that the computer can understand.

- **Optical Disc**: An optical disc drive (ODD) is a device that reads and writes data to an optical disc. It contains a laser, a lens, and a photodetector. The laser reads the data from the disc, and the lens focuses the laser beam onto the disc surface. When a disc is inserted into the drive, the laser reads the data on the disc. The laser beam is focused onto the disc surface, and the photodetector detects the reflection from the disc. The data on the disc is encoded in the form of small pits and lands, which cause variations in

the reflection. To write data to a disc, the ODD uses a process called burning. The laser beam is used to heat a small area of the disc surface, which causes a chemical change in the disc material. This creates a pit in the disc surface, which can be read as a "0" by the ODD. Optical discs have different storage capacities and different read and write speeds. CDs can store up to 700MB of data, while DVDs can store up to 4.7GB, and the latest generation of Blu-ray discs can store up to 100 GB. This makes them useful for storing large amounts of data, such as video and audio files, and they are still used in some areas, such as movies, music, and software distribution.

- **Tape Storage:** Tape storage is a type of data storage that uses magnetic tape to store and retrieve data. The tapes used in tape storage are typically made of a thin strip of plastic coated with a magnetic material, similar to the tapes used in analog cassette tapes. A tape drive is a device that reads and writes data to a tape. It contains a read/write head and a mechanism that moves the tape past the head. The read/write head uses magnetic fields to write data to the tape and read data from the tape. When data is written to a tape, the tape drive uses the read/write head to magnetize small areas of the tape's surface. These magnetized areas represent the binary data that is stored. To read data from a tape, the read/write head detects the magnetic fields on the tape and converts them into binary data that the computer can understand. Tape storage is typically used for the backup and archiving of large amounts of data because it is relatively inexpensive and has high storage capacity.

- **Solid-State Drives (SSD)**: Data is stored on an SSD in small blocks of memory called pages. These pages are grouped together into larger blocks called blocks. To write data to an SSD, the drive first erases a memory block and then writes the new data to the pages within that block. This process is called "block-level write. "When reading data from an SSD, the drive retrieves the requested data from the appropriate page within a block. Because SSDs have no moving parts, they can access data much faster than traditional hard drives, which have to physically locate the correct part of the disk to read the data. Overall, SSDs are faster, more reliable and use less power than traditional hard drives, which makes them ideal for use in laptops, desktop computers, and servers.

- **USB Drives:** USB drives, also known as flash drives or thumb drives, are small, portable storage devices that can be easily connected to a computer. They use non-volatile memory to store data and are typically faster and more reliable than optical discs. They are also relatively inexpensive and easy to use, making them a popular choice for data backup and transfer.

"Modern storage media" refers to the latest advancements in storage technology, which are characterized by their high capacity, high performance, and advanced features. Some examples of modern storage media include:

- **NVMe SSDs**: NVMe (Non-Volatile Memory Express) SSDs are a newer type of SSD that uses the NVMe protocol to communicate with the host system. They offer faster read and write speeds than traditional SSDs, making them suitable for high-performance applications such as gaming, video editing, and data analytics.

- **Cloud Storage**: Cloud storage is a modern storage solution that allows users to store data on remote servers that are connected to the internet. This type of storage is highly scalable, and the user can access their data from anywhere with an internet connection. Cloud storage providers such as Amazon S3, Google Drive, and Microsoft OneDrive are popular among individuals and businesses.

- **Object Storage**: Object storage is a modern storage solution that allows users to store and retrieve data as objects, rather than files or blocks. It is designed for large-scale data storage and is highly scalable and reliable. Object storage solutions such as Amazon S3, Google Cloud Storage, and Microsoft Azure Blob Storage are widely used in big data and cloud computing applications.

- **Memory-based storage**: Memory-based storage, such as Intel Optane and Samsung Z-NAND, is a high-performance storage solution that uses non-volatile memory to store data. This technology provides faster access times and high IOPS, making it ideal for in-memory databases, high-performance computing, and other memory-intensive applications.

3.1 Drawbacks of existing storage media

Current storage mediums such as magnetic tapes, hard drives, and solid-state drives have several limitations compared to DNA.

Limited lifespan: Some storage mediums, such as magnetic tapes and hard disk drives, have a limited lifespan. Over time, these mediums can degrade and lose data, making them unreliable for long-term storage.

Vulnerability to physical damage: Some storage mediums, such as hard disk drives and USB drives, are vulnerable to physical damage, such as scratches and breaks. This can lead to data loss and make the medium unusable.

Vulnerability to cyber threats: Some storage mediums, such as cloud storage and external hard drives, are vulnerable to cyber threats, such as hacking and data breaches. This can lead to data loss or unauthorized access to sensitive information.

Accessibility: Some storage mediums, such as magnetic tapes and DNA storage, are not easily accessible, meaning that the data stored in them can be difficult and/or expensive to retrieve.

Capacity: Some storage mediums, such as memory cards, may have limited storage capacity, making them less suitable for storing large amounts of data.

Environmental impact: Some storage mediums, such as hard disk drives, have a negative impact on the environment due to the use of toxic materials and the energy consumption during the production and disposal of these mediums.

What is DNA and how can it store data?

DNA stands for Deoxyribonucleic Acid, and it is the genetic ingredient that carries the instructions for the development and function of all living organisms on this planet. It is like a blueprint or recipe for how your body is built and works. DNA is made up of a long chain of building blocks called nucleotides, which come in four different types: A, T, C, and G. The order of these nucleotides determines the instructions for the organism. These instructions are passed down from parents to their offspring and determine things like eye color, hair color, and other physical characteristics. In other words, DNA is the code that makes you, you!

A, C, G, and T are the four types of nucleotides that make up the DNA molecule. They are also known as the "**bases**" of DNA.

A (adenine) always pairs with T (thymine)

C (cytosine) always pairs with G (guanine)

The order of these nucleotides, or bases, determines the instructions for the organism. This is because the sequence of these bases' codes for the production of proteins which are responsible for the majority of the functions in living organisms. For example, a specific sequence of bases might code for a certain protein that helps protect the skin from the sun, while another sequence might code for a protein that helps transport oxygen in the blood. By understanding the sequence of these bases, scientists can understand how different organisms develop and function.

The interesting part is, how can it store binary data that our computers can understand?

DNA can store digital data by encoding it into the sequence of nucleotides that make up the DNA molecule. The process of storing digital data in DNA involves the following steps:

Data encoding: The first step is to convert the digital data into a string of binary code, which can be represented using the four nucleotides that make up the DNA code (A, C, G, and T). This process is known as data encoding. The binary code is then translated into a DNA sequence using a mapping algorithm. For example, the binary code "01" can be mapped to the nucleotide "A" and "10" can be mapped to "C" and similarly for other nucleotides as well. Then to ensure the accuracy and reliability of the encoded data, error correction codes are added to the DNA sequence. This allows for the detection and correction of errors that may occur during the encoding process. Now the DNA sequence is ready.

DNA synthesis: Once the DNA sequence is generated, it can be synthesized using a process called PCR (polymerase chain reaction) or by using a DNA synthesizer. PCR is a laboratory technique that amplifies a specific DNA sequence. It involves heating and cooling cycles to separate the DNA strands, and then using an enzyme called polymerase to add nucleotides to the separated strands, creating new DNA molecules with the same sequence as the original. PCR can be used to create multiple copies of a specific DNA sequence Whereas DNA synthesizers or gene synthesizers are specialized machines that can create artificial DNA molecules with a specific sequence. These machines use a process called solid-phase DNA synthesis, which involves adding nucleotides to a solid support (such as a bead or a chip) in a specific order to create the desired DNA sequence. The synthesized DNA is then released from the solid support and can be used for. creating an artificial DNA molecule that contains the encoded digital data.

DNA storage: The synthesized DNA molecules can be stored in a dry and cool environment such as a -20°C freezer. DNA is relatively stable and can potentially last for hundreds or even thousands of years, depending on the storage conditions.

Data retrieval: To retrieve the data, the DNA is sequenced using next-generation sequencing technologies such as Illumina or PacBio. Illumina is a widely used NGS technology that is known for its high throughput, low cost, and high accuracy. It uses a process called "sequencing by synthesis" where fluorescently labeled nucleotides are added to a DNA sample one at a time, and the sequence is read by a camera. This process allows for the rapid sequencing of large amounts of DNA at a relatively low cost. PacBio, on the other hand, uses a process called "single molecule, real-time sequencing" (SMRT) which allows for the direct observation of DNA polymerase activity during the sequencing process. This results in longer read lengths, but at the cost of lower throughput and a higher cost per base. PacBio is often used for applications where long read lengths are needed, such as genome assembly, epigenetic studies, and detection of structural variations. The sequencing process reads the DNA sequence and generates a large amount of data that needs to be processed and analyzed. In summary, both Illumina and PacBio are DNA sequencing technologies, but they have different strengths and are used for different

types of applications. Illumina is fast and accurate, but it reads shorter stretches of DNA, while PacBio reads longer stretches of DNA, but it is slower and more expensive.

Data decoding: The final step is to decode the DNA sequence back into the original binary code and then into the original digital data. DNA data decoding is the process of converting the information stored in a DNA molecule back into its original digital form. This process involves reading the sequence of nucleotides in the DNA and using a reverse mapping algorithm to convert it back into the binary code that was used to encode the data. Imagine you have a message written in a code, in this case, the message is stored in the DNA sequence, and the code is the mapping used to encode the data. To decode the message, you need to use the key (mapping algorithm) used to encode the data, but in reverse. It is like solving a puzzle, you need the same key that was used to lock it, to unlock it. Once the binary code is obtained, it is converted back into the original digital data, which can be a text, image, or any other type of digital information. This process allows the data that has been stored in the DNA molecules to be accessed and used again.

These are the process involved in storing the "base" 2 data in nucleotide "bases". With this method, the scarcity of storage problem in the data era can be tackled in an efficient manner.

Advantages of DNA as a storage device

DNA has several unique properties that make it an attractive candidate for data storage. The most notable property is its storage density, which is incredibly high, with a storage density of about 1 exabyte per gram. This is approximately 10 million times denser than current magnetic tape storage. Additionally, DNA is stable and can potentially last for hundreds or even thousands of years. DNA molecules are also relatively small, which makes them easy to store and transport.

High data density: DNA can store a massive amount of data in a small physical space.

Stability: DNA is highly stable and can last for thousands of years, making it ideal for long-term storage.

Encryption: DNA can be encrypted for added security.

Data retrieval: Data retrieval from DNA is possible by reading the DNA sequence using DNA sequencing techniques.

Future scope of DNA storage

DNA storage is still a relatively new technology and is primarily being used for research and experimental purposes. DNA storage has several potential applications as a device; here are a few examples:

Archival storage: DNA storage could store large amounts of data such as historical documents, art, and cultural heritage for thousands of years.

Backup storage: DNA storage could be used as a backup storage medium for essential data, such as financial records, medical records, and legal documents, which must be kept for long periods of time.

Cold storage: DNA storage is suitable for storing data at low temperatures, and it could be used in cold storage applications, such as deep-sea storage, cryogenic storage, and space storage.

Biomedical research: DNA storage could be used to store large amounts of genetic data and other data relevant to biomedical research.

Secure storage: DNA storage can be encrypted and secured, making it a good option for sensitive data, such as national security data and intellectual property.

Big data: DNA storage could be used to store massive amounts of data generated by the Internet of Things (IoT) devices, social media, and other sources.

Areas of Focus

It is essential to keep in mind that DNA storage is still a developing technology and it is not yet fully commercialized. However, with the rapid advancements in synthetic biology and DNA sequencing techniques, the cost of DNA storage is expected to decrease, and its usage will become more widespread.

There are several areas of focus when it comes to implementing DNA storage as a device:

Data encoding and compression: Developing efficient methods for encoding and compressing data in the DNA sequence is crucial for making DNA storage practical and cost-effective.

DNA synthesis and storage: Advancing the technology for synthesizing DNA at scale and developing methods for storing the synthesized DNA in a stable form is essential for making DNA storage a viable option.

DNA sequencing: Developing accurate and efficient methods for reading the DNA sequence is crucial for retrieving stored data from the DNA.

Error correction: Developing methods for error correction and data protection, as DNA is a biological molecule and it can be affected by different factors such as humidity, temperature, and radiation.

Security: Developing methods for encrypting and securing stored data in DNA is important for protecting sensitive information.

Cost-effective: Lowering the cost of DNA storage is important for making it a viable option for commercial use.

Interoperability: Developing methods for accessing and retrieving data stored in DNA from different platforms and devices.

Standardization: Developing a standardized format for DNA storage to enable interoperability and compatibility across different platforms and devices.

Scalability: Developing methods for scaling up DNA storage to store large amounts of data, and also to make DNA storage a viable option for commercial and industrial use.

Overall, implementing DNA storage as a device requires a multi-disciplinary approach involving researchers from fields such as synthetic biology, computer science, and electrical engineering.

Conclusion

The technology requires new tools and novel applications of existing ones. However, it is possible that in the future, some of the world's most important archives may be stored in a small collection of DNA molecules, making it a more secure, stable, and long-term storage option.

DNA storage has the potential to revolutionize the way we store data. With its high data density, stability, and encryption capabilities, DNA storage could be used for archival storage, backup storage, cold storage, biomedical research, secure storage, and big data.

However, while DNA storage technology currently shows strong promise, significant technical challenges must be overcome to fully integrate it into other technologies and make it a viable option. It can potentially be a future solution for data storage, but it is currently in the initial stages of development and not yet widely used.

As the limits of silicon technology approach, integrating biomolecules into computer design through the use of hybrid silicon and biochemical systems should be seriously considered. Biotechnology has greatly benefited from advancements in silicon technology in the computer industry, and now it is time for computer architects to return the favor by incorporating biomolecules into their designs.

References

[1] https://www.microsoft.com/en-us/research/project/dna-storage/

[2] Illumina and PacBio DNA sequencing data
https://www.sciencedirect.com/science/article/pii/S2352340920306235

[3] DNA for Data Storage and Retrieval : https://fas.org/blogs/sciencepolicy/dna-for-data-storage-and-retrieval/

[4] Storage media : https://www.techtarget.com/searchstorage/definition/storage-medium