Can We Store the Whole World's Data in DNA Storage?

Bingzhe Li[†], Nae Young Song[†], Li Ou[‡], and David H.C. Du[†]

[†]Department of Computer Science and Engineering, University of Minnesota, Twin Cities
[‡]Department of Pediatrics, University of Minnesota, Twin Cities
{lixx1743, song0455, ouxxx045, du}@umn.edu,

Abstract

The total amount of data in the world has been increasing rapidly. However, the increase of data storage capacity is much slower than that of data generation. How to store and archive such a huge amount of data becomes critical and challenging. Synthetic Deoxyribonucleic Acid (DNA) storage is one of the promising candidates with high density and long-term preservation for archival storage systems. The existing works have focused on the achievable feasibility of a small amount of data when using DNA as storage. In this paper, we investigate the scalability and potentials of DNA storage when a huge amount of data, like all available data from the world, is to be stored. First, we investigate the feasible storage capability that can be achieved in a single DNA pool/tube based on current and future technologies. Then, the indexing of DNA storage is explored. Finally, the metadata overhead based on future technology trends is also investigated.

1 Introduction

The amount of the whole world's digital data has been increasing immensely. It is predicted that it will reach 175 Zettabyte (ZB) in 2025 by the International Data Corporation (IDC) [1]. To store data persistently, many types of storage devices have been exploited for decades such as Hard-Disk Drive and Solid-State Drive [2, 3]. With the exponentially increased data amount, the demand for storage capacity to hold them is also rapidly increased. To satisfy such a capacity requirement, new types of storage devices like Shingled Magnetic Recording (SMR) [4, 5] and Interlaced Magnetic Recording (IMR) [6, 7] disks have emerged. In addition to them, magnetic tape [8], the traditional storing media, still has a considerable market portion for archiving data. However, the capacity of existing storage media is not keeping up with the growth rate of digital data created.

To satisfy the demand for storing the increased data amount, DNA as a storage medium has been becoming an attractive choice due to its high spatial density and long durability. The DNA storage can achieve a theoretical density of 455 EB/g [9] and has a long-lasting property of several centuries [10, 11]. These characteristics of DNA storage make it a great candidate for archival storage. Many research studies focused on several research directions including encoding/decoding associated with error correction schemes [11–18], DNA storage systems with microfluidic platforms [19–21], and applications such as database on top of DNA storage [9]. Moreover, several survey papers [22, 23] on DNA storage mainly focused on the technology reviews of how to store data in DNA (*in vivo* or *in vitro*) including the encoding/decoding and synthesis/sequencing processes. In fact, the major focus of these studies was to demonstrate the feasibility of DNA storage with a small amount of digital data. The scalability of DNA storage to hold a vast amount of data is missing in these studies.

In this paper, we first briefly introduce the current technologies to convert digital data to DNA data and the biochemical factors that comprise the DNA storage. Then, according to the current DNA technologies, we build an in-house DNA storage model to investigate the trade-offs between these biochemical parameters/factors for holding a vast amount of digital data. After that, we discuss the scalability issues of DNA storage for storing the whole world's data. We mainly focus on answering these three questions: **I**. What is the capacity of a single isolated DNA pool/tube? And how biochemical parameters affect the capacity? **II**. How to efficiently index the whole world's data in DNA storage? **III**. With future technologies under development, how these improved technologies can affect the capacity of future DNA storage?

2 Background

In this section, we introduce the basic steps for storing and retrieving digital data to and from DNA storage, respectively. We will first discuss various biochemical factors and parameters in these steps. We will especially focus on the factors that affect the scalability of DNA storage.

In DNA, *Nucleotides* are the small molecules that are fun-

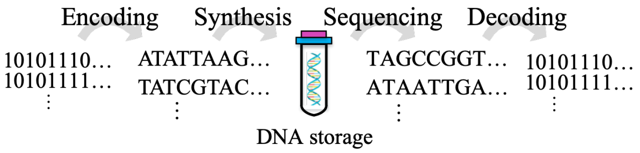


Figure 1: Basic steps of DNA storage.

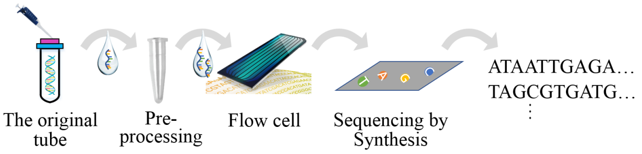


Figure 2: Detailed sequencing process.

damental building blocks of DNA. There are four types of nucleotides: Adenine (A), Cytosine (C), Guanine (G), and Thymine (T). DNA strand is a DNA sequence of several nucleotides. Each DNA nucleotide (nt) refers to a 'base pair (bp)' used in the following sections.

DNA synthesis (write): To store digital data in DNA, we must first transform digital data into DNA sequences. This process is called *encoding*. Theoretically, we can encode two digital bits per one nucleotide by the Quaternary system. We refer to the number of bits per base pair as *encoding density*. Due to the biochemical limitation and internal indexing overhead (i.e., offset and length), most existing works [11, 16, 17, 20, 24–29] have the encoding densities smaller than 2.

We can chemically synthesize a DNA sequence nucleotide by nucleotide as designs [30, 31]. Current technologies can synthesize a DNA strand up to 3,000 bp [32–34]. However, when the length is increased, the error rate happening on each nucleotide bind is also exponentially increased [17, 26, 32, 34, 35]. Due to these reasons, the majority of the existing works for DNA storage use 150~300 bp length of a DNA strand.

To deal with the potential errors existed in DNA strands, using *Error Correction Codes (ECCs)* to ensure the data intactness is inevitable. Thus, some redundant information needs to be added to DNA strands to achieve error correction. In general, in ECC, the more redundancy, the more tolerant of errors it will be. There are two approaches to adding redundancy for error correction purpose. One is adding the ECC within a DNA strand, and the other is, like Redundant Array of Independent Disks (RAID) system, to store redundant data across DNA strands. Both of them achieve the same purpose of recovering errors occurred in DNA strands, but induce a density overhead.

Synthesized DNA strands can be stored in a single isolated pool, and this pool contains a number of different DNA strands. To achieve random access [26] from different DNA strands in the pool, each DNA strand must begin and end with a *primer* to achieve *Polymerase Chain Reaction* (PCR [36]) which is used to duplicate targeted DNA strands with this primer pair. A primer is a short nucleic acid sequence that provides starting and ending points for DNA synthesis. Usually, the length of a primer is between 18~25 bp [37].

We assume each primer has 20 bp such that theoretically, the total number of possible primers is 4^{20}. However, the primer design needs to follow the design guide [37] to avoid some specific conditions such as too high or low portion of GC content and long homopolymers. This is why the number of primers that can be used is much smaller. Being added at the beginning and the end of a DNA strand, the primer pair plays a vital role as a data indicator that enables individual decoding of specific data stored within a single isolated DNA pool. Therefore, the system has to assign a unique primer pair to a set of data chunks that we want to randomly access together. That is, each data chunk stored as a DNA strand, and the whole set of data chunks/DNA strands can be accessed together based on the unique primer pair. The basic steps of DNA processing are depicted in Figure 1. As mentioned before, DNA strands are synthesized and stored in storage such as a tube with a hydrated form like DNA strands dissolved in liquid because the liquid form is easier to handle than a solid form like powder.

DNA sequencing (read): We assume that the data have been synthesized and stored in a single isolated DNA pool/tube, which means that millions of different DNA strands are mixed in one tube. The detailed sequencing process of DNA storage is shown in Figure 2. To read out targeted DNA strands (those synthesized with the same primer pair), the first step is taking one droplet from the DNA tube to amplify the targeted DNA strands via PCR. During the PCR, it takes a specific primer pair as input and duplicates DNA strands with this primer pair. Generally, the PCR process is iterated for multiple cycles to accumulate enough of the targeted DNA strands. After that, a sample of DNA strands is sent to a sequencing machine for the sequencing process. There are several sequencing techniques including Sanger sequencing [38], Next-Generation Sequencing (NGS) [39, 40], and Nanopore Technologies [41]. Different sequencing technologies have different sequencing latencies and error rates. These sequenced DNA strands from the sequencing machine are then decoded into the original digital data. During the decoding, the ECC codes are essential because the sequencing process is also error-prone. Finally, the internal index in DNA strands will help identify the targeted DNA strands.

3 DNA Storage Investigation

In this section, we mainly investigate the scalability of DNA storage based on current technologies which include the capacity of a single DNA pool and the total number of DNA pools needed to store the whole world's data. We assume that the DNA storage implementation is PCR-based random access [26].

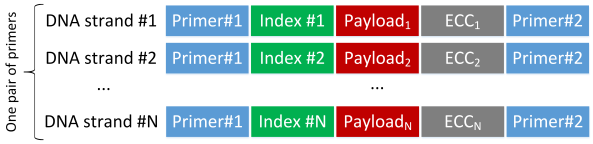


Figure 3: DNA strand format with Parallel Factor (PF).

3.1 DNA Strand Format and Configuration

We look into different factors that can affect the capacity of DNA storage, such as errors, DNA strand length, and encoding density. Basically, one DNA strand contains four fields: primer, index, payload, ECC, and primer as shown in Figure 3. A pair of primers attached to the head and the tail of a DNA strand are used to achieve random access in DNA storage. As explained in the previous section, due to the biochemical limitations such as hairpin, cross/self dimer, and GC Content [26, 37, 42], not all primers can directly be used. Even though the length of a primer is subject to some constraints in real experiments, for the sake of simplicity, we use a generally-used fixed length of 20 bp as the default primer length in this paper. Therefore, according to the existing work [26], the total number of primer pairs in the primer library is 14,000 after the PCR primer design. The maximum number of available primers may be varied based on the data size (average information size associated with a pair of primers) and the contents of stored data. The total number of available primer pairs is decreased generally as the data size increased. As one primer pair can play a role of an indicator of specific data chunks to be read out, we define *Parallel Factor (PF)* referring to the number of different DNA strands associated with one primer pair. In other words, the PF indicates the number of different DNA strands that can be read out with one sequencing process, as described in Figure 3. A larger PF may result in a smaller maximum number of available primer pairs that can be used due to the fact that the encoded DNA content may contain conflicts with some primers.

The payload is the useful information that is encoded from digital binary values to nucleotides. In one DNA strand, the size of the useful data is calculated by (Payload length * Encoding density). Therefore, encoding density is one of the important factors to affect capacity since it indicates how many digital bits can be stored in one bp. According to previous work [11, 16, 17, 20, 24–29], the encoding densities are varied from 0.29 to 1.94 based on different coding schemes.

The ECC field is used to recover any errors induced by the synthesis and sequencing processes. In general, a longer DNA strand will cause a higher error rate, which needs a stronger ability of ECC codes. It means we need to have a higher ECC ratio and thus results in less payload information in DNA strands and lower encoding density. In this paper, we simply set the ECC ratio to 15% as a default based on the paper [26], which uses a Reed–Solomon encoding scheme.

Table 1: The default parameters used in the DNA storage

Total data in world (ZB) in 2020	44 [1]
DNA Strand Length(bp)	300
Primer length (bp)	20
Coding Density	0.29 - 1.94
ECC ratio (Logical redundancy)	15% [26]
Tube size (mL)	1.7
Droplet size (mL)	0.001
Max DNA solubility in liquid (mg/mL)	500 [44–46]
Total number of primer pairs in library	14,000 [26]
Parallel Factor (PF)	$1.55 * 10^6$ [25, 26]
# of DNA copies required by PCR (20 - 25 cycles)	1000

3.2 DNA Storage Modeling

Based on the current technologies, the default parameters are summarized in Table 1. To calculate the capacity of one DNA storage tube, following the read process in Section 2, one droplet from a DNA tube is taken, and the target data in this droplet is read out. Therefore, the droplet should contain all the same data as in the tube. Based on parameters in Table 1 including the maximum solubility of DNA in liquid and DNA calculator [43], we can calculate how many DNA strands that can be dissolved in the droplet. The maximum solubility of DNA in liquid denoted by **Max Solubility**[1] describes how many DNA strands (by weight) can be dissolved in liquid, which is shown as 500 mg/mL[2] in Table 1. Then, if we follow the requirement of PCR and bio restriction such as the number of DNA copies for PCR (1000), PF (10^6), etc., the total number of different DNA strands in one droplet can be computed. At the same time, based on DNA length (300 bp), coding density (maximum 1.94), primer length, and ECC ratio, we can obtain how much useful information can be stored in one DNA strand as shown in Eq. (1). Finally, the total capacity of one DNA tube is computed by the number of different DNA strands multiplying the payload information per DNA strand, which is about **660 GB** per tube. To store the whole world's data 44 ZB (the amount of digital data today), we need more than $7 * 10^{11}$ tubes.

$$L_{DNA} = L_{primer} * 2 + L_{ECC} + L_{payload} + L_{index} \tag{1}$$

3.3 Scalability of Single DNA Pool

Obviously, according to the analysis in Section 3.2, the capacity of 660 GB per tube in DNA storage is too small for archival systems since the current disk drives like SMR drives have reached a storage capacity of 16 TB per drive. Therefore, we need to investigate the possibilities of DNA storage based on future technologies.

According to the discussion in Section 3.1, the DNA storage capacity can be affected by many factors such as DNA

[1] The data in the Max Solubility case can be read out after proper dilution.

[2] The solubility is varied based on temperature, PH value, DNA length, liquid chemical element, etc. [44–46]. We use 500 mg/mL in this paper and the total capacity of a tube is proportional to the solubility.

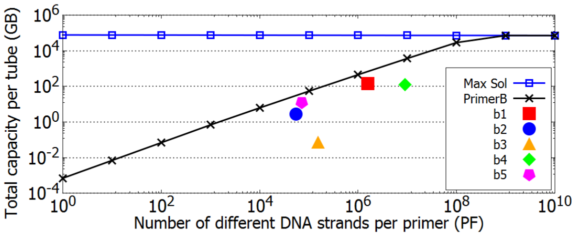


Figure 4: Total capacity in one tube as varying PF.

strand length, ECC ratio, encoding density, and PF. We assume that some of the factors influencing the capacity can be scaled up based on the on-going development of biological technologies. These factors also have an effect on each other when scaling up. We introduce these trade-offs when scaling each factor in the following paragraphs.

Two cases are investigated for the capacity of a single DNA pool. The first one is **Max Solubility** that indicates the theoretical maximum number of DNA strands in one tube so that the capacity calculated by using the max solubility is an upper bound of DNA storage. The other one is **PrimerB** based on a practical assumption and the current technologies in Table 1. The existing works (b1 [26], b2 [20], b3 [24], b4 [25] and b5 [17]) are also plotted in the figures.

PF: As discussed in Section 3.1 and the work [26], when scaling up PF value, the number of available primer pairs is decreased, but the total amount of different DNA strands is increased in one tube. Moreover, the length of internal index is also increased, resulting in less available payload information in a fixed length DNA strand, but more information can be read out in one sequencing. Due to the biochemical restriction (no content can overlap with primers) and synthesis and sequencing practical considerations, the current largest PF value is about $1.55 * 10^6$ [25, 26]. Assuming the PF value can be scaled up with future technologies, we vary PF from 1 to 10^{10}. As shown in Figure 4, the 'Max Solubility' case indicates the theoretical maximum capacity of DNA storage in one tube, which is around 75TB. The capacity of the 'PrimerB' case is increased as the increased PF value and finally saturates to around 75TB at 10^9 due to the maximum solubility of DNA in liquid. The existing studies (b1~b5 [17, 20, 24–26]) use their DNA strand configurations and achieve lower capacities than 'PrimerB' under the same PF values.

DNA strand length: DNA length has a direct impact on the capacity of DNA storage. When increasing DNA length, the payload length (Eq. (1)) is proportionally increased in the ideal case. However, error rates in DNA storage will be higher with a longer DNA length [34, 35]. Therefore, we must use a more powerful ECC to recover errors when using a longer DNA length, resulting in a lower ratio of payload information in the whole DNA length. In future technologies, we simply assume that the synthesis and sequencing technologies will introduce less errors than current technologies and the ECC ratio will keep the same regardless of the increased DNA length. Figure 5 shows the total capacity varying DNA strand length when PF is 10^5 to 10^9. All cases follow a similar trend that the capacity of a tube is increased as increasing the DNA strand length. A longer DNA strand means that each strand contains more payload information, and its size has 113 times difference between lengths of 100 and 3,000. Since a longer DNA strand length decreases the number of DNA strands in one droplet due to the maximum solubility, the 'Max Solubility' only obtains around 2.3x increase from the DNA length of 100 to 3,000. The 'PrimerB' achieves about 2.7x-51x increase because the number of available primers is decreased as the DNA length increases.

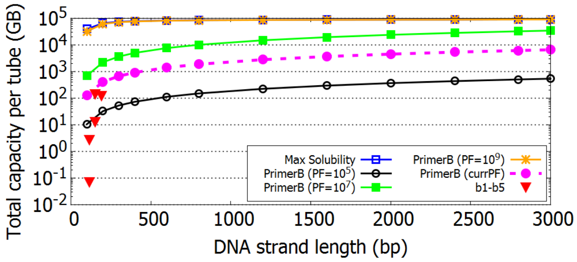


Figure 5: Total capacity with varying DNA strand length. 'currPF' refers to the current technology with PF=$1.55 * 10^6$.

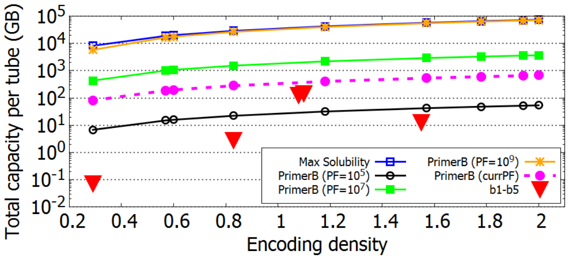


Figure 6: Total capacity with varying encoding logical density.

Encoding density: Encoding density has an influence on other factors. With a higher encoding density, one DNA base pair can store more digital information. Moreover, under the same error correction ability, ECC codes with higher encoding density need fewer base pairs, resulting in a longer payload length. However, due to the biochemical limitation, a higher encoding density causes a higher error rate.

We investigate the total capacity of one tube by varying the density from 0.29 to 2. As shown in Figure 6, as encoding density increases, the capacity increases about 7.9x-12.3x, which is higher than the density increase (6.9x from 0.29 to 2). The reason is that a higher density not only increases the payload information but also shrinks the length of the internal index, thus further improves the capacity.

Maximum capacity: If we take the aggressive future technology configuration (coding density 2, $PF = 10^9$, and DNA strand length of 3,000), the total capacity of a tube is about 90 TB, and we need $5 * 10^8$ tubes to store the current whole world's data (44 ZB), which is about 1400x reduction com-

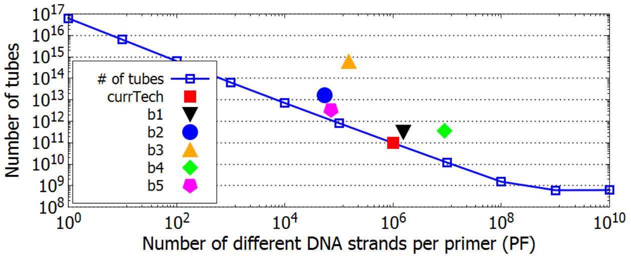
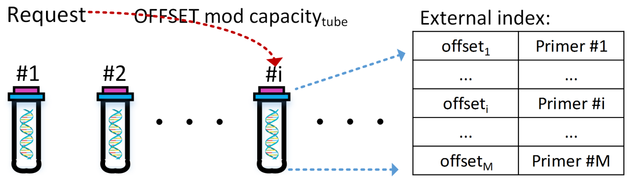
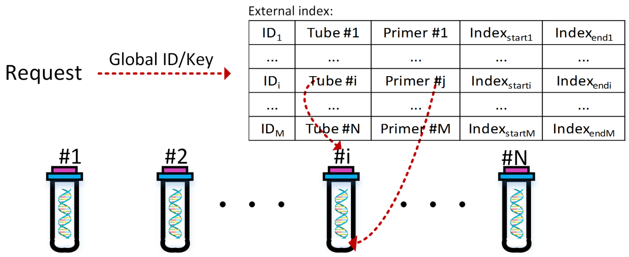


Figure 7: Total number of physical tubes to store the whole world's data.



offset_1	Primer #1
...	...
offset_i	Primer #i
...	...
offset_M	Primer #M

Figure 8: Indexing scheme for the Block-based scenario.



ID_1	Tube #1	Primer #1	$\text{Index}_{\text{start1}}$	$\text{Index}_{\text{end1}}$
...	...	...	...	...
ID_i	Tube #i	Primer #j	$\text{Index}_{\text{starti}}$	$\text{Index}_{\text{endi}}$
...	...	...	...	...
ID_M	Tube #N	Primer #M	$\text{Index}_{\text{startM}}$	$\text{Index}_{\text{endM}}$

Figure 9: Indexing scheme for the Object-based scenario.

pared to that with the current technology. To store future world's data (175 ZB in 2025 [1]), we need about $2 * 10^9$ tubes.

4 Indexing of DNA Storage

Based on the discussion in Section 3.3, more than 10^{11} tubes are needed for storing the whole world's data, as shown in Figure 7. Therefore, how to index stored data can be a critical issue. In this section, we investigate two scenarios of DNA storage indexing by using simple schemes to discuss the indexing overhead of DNA storage.

First, we investigate two scenarios. One is to use the DNA storage as a block I/O device, which means that we capture and store data based on its offset, denoted by **Block-based**. In this scenario, we can recover all digital data by retrieving all data from DNA storage. The other one is to use DNA storage as an object storage device, which means that we can search and retrieve data based on its Global Unique ID (GUID) or its key, denoted by **Object-based**.

For the Block-based implementation, a straightforward two-tier indexing/mapping scheme is used to include both internal and external indexes. Since a primer pair in a tube can be adapted to many different DNA strands, similar to [26], the internal indexing is used to identify one or several target DNA strands depending on the index field of a DNA strand. First, we directly compute the tube number and then find out the target primer pairs according to the external indexing table, as shown in Figure 8. After that, the DNA strands associated with the target primer pairs are amplified and decoded. Finally, based on the internal indexing, we can decode DNA strands and read the target data out.

For the Object-based implementation, as seen in Figure 9, there is a global external indexing table to indicate the tube number and the primer pair of the target GUID/key. Once getting this information, the sequencing process is the same as that of the Block-based version to read the data out. Finally, depending on internal indexing (start and end index information), we can find the data of the target GUID or the key.

To investigate the overhead of indexing metadata for both Block- and Object-based implementations, we scale different factors considering the future biological technologies. As discussed in Section 3.3, the DNA strand length and PF value have a significant influence on the DNA storage capacity. So, we scale the PF from 1 to 10^{10} and DNA strand length from 100 bp to 5000 bp. We use the maximum encoding density and assume 15% ECC ratio is adequate in this investigation. The Object-based implementation uses the average object sizes of 4KB, 16KB, 1MB, and 4MB.

For the overhead of the indexing metadata for both implementations, Figure 10 and Figure 11 indicate that the Block-based scheme requires about 1.8TB to 25.8ZB indexing metadata and the Object-based scheme needs around 0.189ZB, 46.7EB, 700PB and 42.3PB with different average sizes of data objects. We can find that the Block-based implementation is very sensitive to the PF and DNA strand length. Therefore, the total indexing metadata is proportional to the number of tubes, and is equal to the number of tubes multiplied by the single external indexing table size. On the other hand, the Object-based implementation achieves similar indexing overhead if the number of GUIDs/keys is the same. However, with varying the average object sizes, the indexing overhead has a significant difference. Note that we assume the same total size.

5 Conclusion

In this paper, we investigate the scalability of DNA storage based on current technologies and discuss the bottlenecks when storing the whole world's data in DNA storage. We also identify a few future research issues. We evaluate the potential of DNA storage as a good and reliable alternative for long-term data archive storage in the future. As research of DNA storage progresses, to store the whole world's data, the future

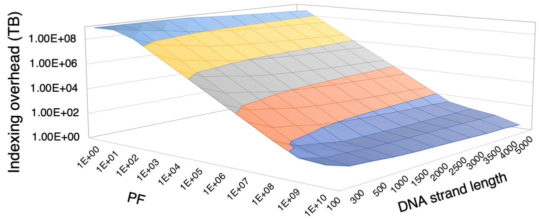


Figure 10: Overall indexing overhead for Block based implementation with future technologies.

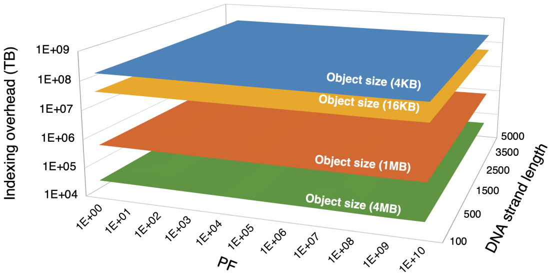


Figure 11: Overall indexing overhead for Object based implementation with future technologies.

DNA storage system can reduce the number of tubes about 1400x compared to that with current technologies. Finally, two storage systems, including Block-based and Object-based systems and their metadata overhead based on the trend of future technologies, are investigated.

6 Discussion Topics

DNA storage is a promising storage device, but many issues still need to be dissolved. We point out several main concerns below.

What will the DNA storage look like in the future? How will the whole system be implemented? The PurpleDrop and puddle [19, 47] provide a good example of the whole system end-to-end. However, there might be some potential issues raised. For example, for each read, we need to consume one droplet and will wash it away after reading the data. As a result, the DNA storage will have a limited number of reads, which is similar to the limited P/E cycles of SSDs. When do we need to replenish the DNA pool? And by how much? Reducing the number of read operations seems to be important.

Moreover, there is still a possibility of insertion or update operations even though it is used for archival storage. This involves a chemical reaction to convert digital values to DNA nucleotides and manipulating DNA strands. Can we do in-place update directly in a DNA pool, or do we need to do out-of-place updates? Since those chemical reactions need to add some extra nucleotides in the liquid, it potentially creates some noise. If we pursue 'real-time' writes and updates, how to deal with this noise? If we do some post-operation to remove or purify the target DNA strands, how do we tolerate those long update/write latencies?

The long latency of write and read of DNA storage may be a critical issue of using DNA storage as well. Based on the current synthesizing and sequencing technologies, the latencies of write and read may reach to hours, which is not tolerable even for archival storage systems. From the storage system research point of view, we have used tiered storage for the combination of fast and slow storage devices. However, whether or not we can successfully adapt DNA storage into our traditional tiered storage systems remains to be investigated. Unlike other traditional systems, any tiered storage including DNA storage should tolerate much more computing or access time. Therefore, this property of DNA storage may provide an opportunity to propose a new type of tiered storage system specially designed for DNA storage.

Acknowledgment

This work was partially supported by NSF I/UCRC Center Research in Intelligent Storage and the following NSF awards 1439622, 1525617, 1536447, and 1812537.

References

[1] David Reinsel, John Gantz, and John Rydning. The digitization of the world from edge to core. *IDC White Paper*, 2018.

[2] Jae-Duk Lee, Sung-Hoi Hur, and Jung-Dal Choi. Effects of floating-gate interference on nand flash memory cell operation. *IEEE Electron Device Letters*, 23(5):264–266, 2002.

[3] Bingzhe Li, Chunhua Deng, Jinfeng Yang, David Lilja, Bo Yuan, and David Du. Haml-ssd: A hardware accelerated hotness-aware machine learning based ssd management. In *38th IEEE/ACM International Conference on Computer-Aided Design, ICCAD 2019*, page 8942140. Institute of Electrical and Electronics Engineers Inc., 2019.

[4] Ahmed Amer, JoAnne Holliday, Darrell DE Long, Ethan L Miller, Jehan-François Pâris, and Thomas Schwarz. Data management and layout for shingled magnetic recording. *IEEE Transactions on Magnetics*, 47(10):3691–3697, 2011.

[5] Fenggang Wu, Bingzhe Li, Zhichao Cao, Baoquan Zhang, Ming-Hong Yang, Hao Wen, and David HC Du. Zonealloy: Elastic data and space management for hybrid {SMR} drives. In *11th {USENIX} Workshop on Hot Topics in Storage and File Systems (HotStorage 19)*, 2019.

[6] Mohammad Hossein Hajkazemi, Ajay Narayan Kulkarni, Peter Desnoyers, and Timothy R Feldman. Track-based translation layers for interlaced magnetic recording. In *2019 {USENIX} Annual Technical Conference ({USENIX}{ATC} 19)*, pages 821–832, 2019.

[7] Fenggang Wu, Baoquan Zhang, Zhichao Cao, Hao Wen, Bingzhe Li, Jim Diehl, Guohua Wang, and David HC Du. Data management design for interlaced magnetic recording. In *10th USENIX Workshop on Hot Topics in Storage and File Systems (HotStorage 18)*, 2018.

[8] Lto ultrium roadmap. http://www.ltoultrium.com/lto-ultrium-roadmap/.

[9] Raja Appuswamy, Kevin Le Brigand, Pascal Barbry, Marc Antonini, Olivier Madderson, Paul Freemont, James McDonald, and Thomas Heinis. Oligoarchive: Using dna in the dbms storage hierarchy. In *CIDR*, 2019.

[10] Morten E Allentoft, Matthew Collins, David Harker, James Haile, Charlotte L Oskam, Marie L Hale, Paula F Campos, Jose A Samaniego, M Thomas P Gilbert, Eske Willerslev, et al. The half-life of dna in bone: measuring decay kinetics in 158 dated fossils. *Proceedings of the Royal Society B: Biological Sciences*, 279(1748):4724–4733, 2012.

[11] Robert N Grass, Reinhard Heckel, Michela Puddu, Daniela Paunescu, and Wendelin J Stark. Robust chemical preservation of digital information on dna in silica with error-correcting codes. *Angewandte Chemie International Edition*, 54(8):2552–2555, 2015.

[12] Michael Luby. Lt codes. In *The 43rd Annual IEEE Symposium on Foundations of Computer Science, 2002. Proceedings.*, pages 271–280. IEEE, 2002.

[13] David JC MacKay. Fountain codes. *IEE Proceedings-Communications*, 152(6):1062–1068, 2005.

[14] Yixin Wang, Md Noor-A-Rahim, Jingyun Zhang, Erry Gunawan, Yong Liang Guan, and Chueh Loo Poh. High capacity dna data storage with variable-length oligonucleotides using repeat accumulate code and hybrid mapping. *Journal of Biological Engineering*, 13(1):89, 2019.

[15] Daniel G Gibson, John I Glass, Carole Lartigue, Vladimir N Noskov, Ray-Yuan Chuang, Mikkel A Algire, Gwynedd A Benders, Michael G Montague, Li Ma, Monzia M Moodie, et al. Creation of a bacterial cell controlled by a chemically synthesized genome. *science*, 329(5987):52–56, 2010.

[16] James Bornholt, Randolph Lopez, Douglas M Carmean, Luis Ceze, Georg Seelig, and Karin Strauss. A dna-based archival storage system. In *Proceedings of the Twenty-First International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 637–649, 2016.

[17] Yaniv Erlich and Dina Zielinski. Dna fountain enables a robust and efficient storage architecture. *Science*, 355(6328):950–954, 2017.

[18] Reinhard Heckel, Ilan Shomorony, Kannan Ramchandran, and NC David. Fundamental limits of dna storage systems. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3130–3134. IEEE, 2017.

[19] Max Willsey, Ashley P Stephenson, Chris Takahashi, Pranav Vaid, Bichlien H Nguyen, Michal Piszczek, Christine Betts, Sharon Newman, Sarang Joshi, Karin Strauss, et al. Puddle: A dynamic, error-correcting, full-stack microfluidics platform. In *Proceedings of the Twenty-Fourth International Conference on Architectural Support for Programming Languages and Operating Systems*, pages 183–197, 2019.

[20] George M Church, Yuan Gao, and Sriram Kosuri. Next-generation digital information storage in dna. *Science*, 337(6102):1628–1628, 2012.

[21] Christopher N Takahashi, Bichlien H Nguyen, Karin Strauss, and Luis Ceze. Demonstration of end-to-end automation of dna data storage. *Scientific reports*, 9(1):1–5, 2019.

[22] Luis Ceze, Jeff Nivala, and Karin Strauss. Molecular digital data storage using dna. *Nature Reviews Genetics*, 20(8):456–466, 2019.

[23] Yiming Dong, Fajia Sun, Zhi Ping, Qi Ouyang, and Long Qian. Dna storage: research landscape and future prospects. *National Science Review*, 2020.

[24] Nick Goldman, Paul Bertone, Siyuan Chen, Christophe Dessimoz, Emily M LeProust, Botond Sipos, and Ewan Birney. Towards practical, high-capacity, low-maintenance information storage in synthesized dna. *Nature*, 494(7435):77–80, 2013.

[25] Meinolf Blawat, Klaus Gaedke, Ingo Huetter, Xiao-Ming Chen, Brian Turczyk, Samuel Inverso, Benjamin W Pruitt, and George M Church. Forward error correction for dna data storage. *Procedia Computer Science*, 80:1011–1022, 2016.

[26] Lee Organick, Siena Dumas Ang, Yuan-Jyue Chen, Randolph Lopez, Sergey Yekhanin, Konstantin Makarychev, Miklos Z Racz, Govinda Kamath, Parikshit Gopalan, Bichlien Nguyen, et al. Random access in large-scale dna data storage. *Nature biotechnology*, 36(3):242, 2018.

[27] Leon Anavy, Inbal Vaknin, Orna Atar, Roee Amit, and Zohar Yakhini. Improved dna based storage capacity and fidelity using composite dna letters. *bioRxiv*, page 433524, 2018.

[28] Yeongjae Choi, Taehoon Ryu, Amos Lee, Hansol Choi, Hansaem Lee, Jaejun Park, Suk-Heung Song, Seoju Kim, Hyeli Kim, Wook Park, et al. Addition of degenerate bases to dna-based data storage for increased information capacity. *bioRxiv*, page 367052, 2018.

[29] Henry H Lee, Reza Kalhor, Naveen Goela, Jean Bolot, and George M Church. Enzymatic dna synthesis for digital information storage. *bioRxiv*, page 348987, 2018.

[30] Sriram Kosuri and George M Church. Large-scale de novo dna synthesis: technologies and applications. *Nature methods*, 11(5):499, 2014.

[31] Mark Douglas Matteucci and M Ho Caruthers. Synthesis of deoxyoligonucleotides on a polymer support. *Journal of the American Chemical Society*, 103(11):3185–3191, 1981.

[32] SM Hossein Tabatabaei Yazdi, Yongbo Yuan, Jian Ma, Huimin Zhao, and Olgica Milenkovic. A rewritable, random-access dna-based storage system. *Scientific reports*, 5:14138, 2015.

[33] S. M. Hossein TabatabaeiYazdi, Ryan Gabrys, and Olgica Milenkovic. Portable and Error-Free DNA-Based Data Storage. *Scientific Reports*, 7(1):1–6, 2017.

[34] Evaluation of linear synthetic dna fragments from separate suppliers. https://www.thermofisher.com/content/dam/LifeTech/global/life-sciences/Cloning/gene-synthesis/PDF/GeneArt%20Strings%20compared%20to%20gBlocks.pdf.

[35] Peter Richterich. Estimation of errors in 'Raw' DNA sequences: A validation study. *Genome Research*, 8(3):251–259, 1998.

[36] John M. S. Bartlett and David Stirling. *A Short History of the Polymerase Chain Reaction*, pages 3–6. Humana Press, Totowa, NJ, 2003.

[37] Pcr primer design guidelines. http://www.premierbiosoft.com/tech_notes/PCR_Primer_Design.html.

[38] Fred Sanger and Alan R Coulson. A rapid method for determining sequences in dna by primed synthesis with dna polymerase. *Journal of molecular biology*, 94(3):441–448, 1975.

[39] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18, 2008.

[40] Wilhelm J. Ansorge. Next-generation DNA sequencing techniques. *New Biotechnology*, 25(4):195–203, 2009.

[41] Oxford nanopore technologies. https://nanoporetech.com/.

[42] Kyle J Tomek, Kevin Volkel, Alexander Simpson, Austin G Hass, Elaine W Indermaur, James M Tuck, and Albert J Keung. Driving the scalability of dna-based information storage systems. *ACS synthetic biology*, 8(6):1241–1248, 2019.

[43] Calculator for determining the number of copies of a template. https://cels.uri.edu/gsc/cndna.html.

[44] Ashok Garai, Debostuti Ghoshdastidar, Sanjib Senapati, and Prabal K Maiti. Ionic liquids make dna rigid. *The Journal of chemical physics*, 149(4):045104, 2018.

[45] Kristi KH Stanlis and J Richard McIntosh. Single-strand dna aptamers as probes for protein localization in cells. *Journal of Histochemistry & Cytochemistry*, 51(6):797–808, 2003.

[46] Young-Wan Kwon, Chang Hoon Lee, Dong-Hoon Choi, and Jung-Il Jin. Materials science of dna. *Journal of Materials Chemistry*, 19(10):1353–1380, 2009.

[47] Sharon Newman, Ashley P Stephenson, Max Willsey, Bichlien H Nguyen, Christopher N Takahashi, Karin Strauss, and Luis Ceze. High density dna data storage library via dehydration with digital microfluidic retrieval. *Nature communications*, 10(1):1–6, 2019.